Tampere University

KIRSI VARPA

On Knowledge Discovery Experimented with Otoneurological Data

Tampere University Dissertations 31

Tampere University Dissertations 31

KIRSI VARPA

On Knowledge Discovery Experimented with Otoneurological Data

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty Council of the Faculty of Information Technology
and Communication Sciences
of the Tampere University,
for public discussion in the auditorium B1097
of the Pinni B building, Kanslerinrinne 1, Tampere,
on 08.03.2019, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos	Professor, Ph.D. Martti Juhola Tampere University Finland	
Supervisor	University Lecturer, Ph.D. Kati Iltanen Tampere University Finland	
Pre-examiners	Senior Lecturer, Ph.D. Jagdish Patra Swinburne University of Technology Australia	Professor, Ph.D. Iren Valova University of Massachusetts Dartmouth United States
Opponent	Senior University Lecturer, D.Sc. Jaakko Hollmén Aalto University Finland	

Cover design: Roihu Inc.

ISBN 978-952-03-1026-4 (print)
ISBN 978-952-03-1027-1 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1027-1

PunaMusta Oy – Yliopistopaino
Tampere 2019

ACKNOWLEDGEMENTS

At first, I wish to express my gratitude to my supervisors, professor Martti Juhola and university lecturer Kati Iltanen. They patiently guided me through this process even though there were difficult times between. Martti did not give up the hope during the years and here we are - finally. Kati, you continuously supported me and your enthusiasm encouraged me many times to push forward.

Making science is not possible without good colleagues. Many times it has been necessary to speak in order to understand what to do. I want to express my gratitude again to Kati Iltanen and to my friend Turkka Näppilä who has always been there during the years at University of Tampere. Perhaps I would not be here without you. Our discussions encouraged me to continue (and vice versa). I want to thank all the colleagues in the Data Analysis Research Group (DARG), especially Henry Joutsijoki and Jyrki Rasku and former office mates, Pekka Niemenlehto and Jyri Saarikoski.

I want to thank the Faculty of Natural Sciences (former School of Information Sciences, former Department of Computer Sciences) for its support to this dissertation. It was nice to see that even if I was "just a grant researcher", I was counted in like a real employee and given a place where to continue my work even after the doctoral school position. So, big thanks to the whole LUO (SIS/CS) personnel and the administration people.

This research would not be possible without the financial support of doctoral programme and several foundations. Therefore, I want to express my gratitude to the departed Tampere Doctoral Programme in Information Science and Engineering (TISE) for the researcher position and thank Finnish Concordia Fund, 'Finnish Cultural Foundation, Päijät-Häme Regional Fund', Onni and Hilja Tuovinen Foundation, Oskar Öflund Foundation, The Ella and Georg Ehrnrooth Foundation, The Scientific Foundation of the City of Tampere, The University of Tampere and University of Tampere Foundation for their grants. Without your support this would had never happen. Thank you!

I want to thank the otoneurological experts, docent, MD Erna Kentala, MD Sari Mykkänen and professor, MD Ilmari Pyykkö for their help in collecting the otoneurological data and medical advice. Thanks goes also to Yrjö Auramo and Matti

J. Tapani due to their previous work with the decision support system ONE. I acknowledge also support of the Finnish IT Center for Science (CSC) that gave its supercomputer resources to use during the genetic algorithm research.

I want to thank the pre-reviewers of my thesis manuscript, senior lecturer Jagdish Patra and professor Iren Valova. In addition, I want to thank senior university lecturer Jaakko Hollmén who has agreed to be my opponent in the public defence.

Sometimes something else than work is also needed. Piia Reku, thank you for listening and supporting me during the years. I want to express my gratitude also to Stina Boedeker who offered me excellent relax breaks during SIS years and boldly joined me in new experiences. She was also a great support during the last steps. Thanks go also to families Nihtilä, Näppilä and Siltanen who have given me opportunities to think totally something else.

Last but not least, I want to thank my family. The path from the beginning to the end was not easy but we managed to get it through. Life wins, eventually. The lost ones will stay in our hearts and memories. Special thanks to my godchildren, niece Menni and nephew Peetu who has bring joy and happiness in our lives. Never stop enjoying and wondering things in life.

Tampere, November 2018
Kirsi Varpa

ABSTRACT

Diagnosis of otoneurological diseases can be challenging due to similar kind of and overlapping symptoms that can also vary over time. Thus, systems to support and aid diagnosis of vertiginous patients are considered beneficial. This study continues refinement of an otoneurological decision support system ONE and its knowledge base. The aim of the study is to improve the classification accuracy of nine otoneurological diseases in real world situations by applying machine learning methods to knowledge discovery in the otoneurological domain.

The phases of the dissertation is divided into three parts: fitness value formation for attribute values, attribute weighting and classification task redefinition. The first phase concentrates on the knowledge update of the ONE with the domain experts and on the knowledge discovery method that forms the fitness values for the values of the attributes. The knowledge base of the ONE needed update due to changes made to data collection questionnaire. The effect of machine learnt fitness values on classification are examined and classification results are compared to the knowledge set by the experts and their combinations. Classification performance of nearest pattern method of the ONE is compared to *k*-nearest neighbour method (*k*-NN) and Naïve Bayes (NB). The second phase concentrates on the attribute weighting. Scatter method and instance-based learning algorithms IB4 and IB1w are applied in the attribute weighting. These machine learnt attribute weights in addition to the weights defined by the domain experts and equal weighting are tested with the classification method of the ONE and attribute weighted *k*-NN with One-vs-All classifiers (*wk*-NN OVA). Genetic algorithm (GA) approach is examined in the attribute weighting. The machine learnt weight sets are utilized as a starting point with the GA. Populations (the weight sets) are evaluated with the classification method of the ONE, the *wk*-NN OVA and attribute weighted *k*-NN using neighbour's class-based attribute weighting (*cwk*-NN). In the third phase, the effect of the classification task redefinition is examined. The multi-class classification task is separated into several binary classification tasks. The binary classification is studied without attribute weighting with the *k*-NN and support vector machines (SVM).

Keywords: Machine learning, knowledge discovery, decision support system, attribute weighting, otoneurological data, vertigo.

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

Abbreviation	Description
ACC	(Classification) Accuracy
ANE	Acoustic Neurinoma (aka Vestibular Schwannoma)
BPV	Benign Positional Vertigo
BRV	Benign Recurrent Vertigo
CL	Central Lesion
CV	Cross-Validation
DSS	Decision Support System
ES	Expert System
GA	Genetic Algorithm
HUCH	Helsinki University Central Hospital
HVDM	Heterogeneous Value Difference Metric
IB	Instance-Based
IDSS	Intelligent Decision Support System
k-NN	*k*-Nearest Neighbour
KB	Knowledge Base
KD	Knowledge Discovery
MEN	Menière's Disease
ML	Machine Learning
MLP	Multiayer Perceptron
NB	Naïve Bayes
ONE	OtoNeurological Expert system
OVA	One-vs-All other
OVO	One-vs-One
PCA	Principal Component Analysis
RBF	Radial Basis Function
SUD	Sudden Deafness
SVM	Support Vector Machine
TAUH	Tampere University Hospital
TPR	True Positive Rate

TRA	Traumatic Vertigo
VES	Vestibulopatia
VDM	Value Difference Metric
VNE	Vestibular Neuritis

ORIGINAL PUBLICATIONS

This dissertation is based on the following five publications. In the text, they are referred to by their Roman numerals.

Publication I Varpa K, Iltanen K, Juhola M, Kentala E and Pyykkö I. Refinement of the otoneurological decision support system and its knowledge acquisition process. In: Engelbrecht R and Hasman A (eds.), *European Notes in Medical Informatics: Ubiquity: Technologies for Better Health in Aging Societies, vol. II, no. 2, 2006. Proceedings of the 20th International Congress of the European Federation for Medical Informatics (MIE 2006)*, Maastricht, Netherlands, 2006, pp. 197–202.

Publication II Varpa K, Iltanen K and Juhola M. Machine learning method for knowledge discovery experimented with otoneurological data. *Computer Methods and Programs in Biomedicine* 91(2), 2008, pp. 154–164. https://doi.org/10.1016/j.cmpb.2008.03.003

Publication III Varpa K, Iltanen K, Siermala M and Juhola M. Attribute weighting with Scatter and instance-based learning methods evaluated with otoneurological data. *International Journal of Data Science* 2(3), 2017, pp. 173–204. https://doi.org/10.1504/IJDS.2017.10007392

Publication IV Varpa K, Iltanen K and Juhola M. Genetic algorithm based approach in attribute weighting for a medical data set. *Journal of Computational Medicine* 2014, 2014, pp. 1–11. https://doi.org/10.1155/2014/526801

Publication V Varpa K, Joutsijoki H, Iltanen K and Juhola M. Applying one-vs-one and one-vs-all classifiers in *k*-nearest neighbour method and support vector machines to an otoneurological multi-class problem. In: Moen A *et al.* (eds.), *Studies in Health Technology and Informatics vol. 169, 2011: User Centred Networked Health Care – Proceedings of 23rd International Conference of the European Federation for Medical Informatics (MIE 2011)*, Oslo, Norway, IOS Press, 2011, pp. 579–583. https://doi.org/10.3233/978-1-60750-806-9-579

1 INTRODUCTION

Domain knowledge is in a key role in decision making. Decisions can concern, for example, diagnosis and treatment of patients, selection of certain products into production, support of students in their studies or ways to enhance customer satisfaction. Whenever there is enough data collected from the domain, it is possible to utilize machine learning (ML) methods [Mitchell, 1997] in finding regularities, rules and/or patterns from the data that can support the decision making. For example, medical diagnostic knowledge can be derived from patients' medical history automatically with machine learning methods and then utilized to assist physician in the diagnosis of new patients in order to improve the diagnostic accuracy, reliability and/or speed [Kononenko *et al.*, 1998].

Knowledge-based systems are based on the knowledge obtained from the domain [Waterman, 1986]. The development of the knowledge-based systems was started already in the mid-1960s [Turban, 1993]. One of the first medical knowledge-based systems was MYCIN [Shortliffe, 1976], which aim was to assist physicians with clinical decisions concerning the selection of appropriate therapy for patients with infectious blood diseases. Decision support systems (DSS) are knowledge-based systems that are utilized to assist decision makers in complex decision making and problem solving [Turban, 1993; Shim *et al.*, 2002] whereas as expert systems (ES) are referred knowledge-based systems that imitate the reasoning process of human experts and use domain-specific knowledge in solving specific problems in a bounded domain of expertise [Turban, 1993; Liou, 1998; Metaxiotis and Samouilidis, 2000]. Decision support systems utilizing artificial intelligence techniques, like machine learning, to enhance support for the decision maker are nowadays referred as intelligent decision support systems (IDSS) [Phillips-Wren, 2013]. Medical diagnosis systems applying machine learning are meant to be helpful tools that can improve the physicians' decision making, not to replace them [Kononenko *et al.*, 1998].

The development of an OtoNeurological Expert system (ONE) to support diagnosis of diseases involving vertigo and to work as an educational tool for medical students was started in the 1990s [Kentala, 1996b; Kentala *et al.*, 1996; Auramo, 1999]. In this dissertation and in the publications, the ONE is referred as

otoneurological decision support system instead of the expert system due to its main purpose to support decision making of the domain experts. Before the ONE, there existed two expert systems for vertiginous patients, Vertigo [Schmid *et al.*, 1987] and Carnisel [Gavilán *et al.*, 1990]. The Vertigo was a rule-based expert system applying Bayesian approach that was meant to be used in a clinical and educational environment as a diagnostic aid for the classification and diagnosis of vestibular disorders [Schmid *et al.*, 1987]. It could classify 32 vertiginous syndromes. The Carnisel was a rule-based expert system made with Prolog: it contained rules and metarules to use in inference but it needed complete information in order to work, it gave only the certified diagnosis [Gavilán *et al.*, 1990]. Recently, an EU-funded EMBalance project has been started. Its aim is to develop a web-based diagnostic decision support system to provide decision support for general practitioners and experts in the diagnosis of 12 balance disorders [Exarchos *et al.*, 2016] and advise on efficient treatment of the patient [Rammazzo *et al.* 2016]. Another aim is to provide a recommendation tool able to guide physicians in requesting the appropriate information of the patient for reaching the diagnosis. The EMBalance decision support system is based on decision trees (C4.5 algorithm [Quinlan, 1993] and ADABOOST [Freund and Schapire, 1997]). It consists of two different modules, one for expert use and another for general practitioners. The EMBalance DSS was evaluated with 985 patients from 12 different balance disorder classes: The classification accuracies varied with diseases from 59.3% to 89.8% with the general practitioner module and from 74.3% to 92.1% with the expert module [Exarchos *et al.*, 2016]. In addition, an intelligent clinical decision making system to support diagnostics of 22 vertigo diseases have been developed [Dong *et al.*, 2014]. The system is based on dynamic uncertain causality graphs. Its total classification accuracy with 60 vertigo cases from 18 disease classes with incomplete data was 81.7% and with complete data 88.3% whereas the physicians classified correctly 53.3%–73.3% of the cases with incomplete data and 70.0%–88.3% with the complete data [Dong *et al.*, 2014].

In the otoneurological domain, several machine learning methods have been applied in classification and knowledge formation, for example, Bayesian probabilistic models [Miettinen and Juhola, 2010], decision tree induction [Viikki, 2002; Exarchos *et al.*, 2016], dynamic uncertain causality graphs [Dong *et al.*, 2014], fuzzy rules induction [Boháĉik and Juhola, 2008], genetic algorithms to discover diagnostic rules [Laurikkala *et al.*, 2001], linear discriminant analysis and *k*-means clustering [Juhola, 2008], neural networks [Juhola *et al.*, 2001; Siermala and Juhola,

2006; Autio *et al.*, 2007] and support vector machines [Joutsijoki *et al.*, 2013]. In addition, statistical methods have been experimented in attribute weighting [Syed, 2014].

The inference of the ONE was evaluated by comparing its inference results to the Vertigo with 365 cases [Auramo and Juhola, 1995]. The Vertigo could give valid diagnosis for 32.3% of the cases whereas the ONE gave valid diagnoses to 78.9% of the cases, thus, demonstrating and having better reasoning result than the Vertigo. The decision making ability of the ONE was also compared to the diagnoses of six physicians of otolaryngology with 23 cases [Kentala *et al.*, 1998]. With the same information about the patients in use, the ONE diagnosed 65.2% of the cases correctly, whereas, the physicians diagnosed on average 54% of the cases correctly. Typically, machine learnt diagnostic rules outperform slightly the diagnostic accuracy of physicians when having exactly the same information in use [Kononenko *et al.*, 1998]. When the physicians had the patients' full medical history in use, they diagnosed about 69% of the cases correctly. It was noticed that the incorrectly classified test cases had confounding symptoms (for example, noise-induced hearing loss or symptoms caused by another disease), which affected the classification results also with the ONE [Kentala *et al.*, 1998]. Thus, the knowledge of the ONE was shown to need further refinement in order to work properly in real world situations with the cases having confounding symptoms.

The suitability of the ML methods for refining and expanding the knowledge for the six largest vertigo disease classes in the database of the ONE was examined in the dissertation of Viikki [Viikki, 2002]. Special attention in Viikki's study was given to the acquisition of diagnostic knowledge and data pre-processing, especially the feature subset selection. The main ML method applied in Viikki's study was decision tree induction [Quinlan, 1993]. Viikki's research showed that the knowledge acquisition process can be eased with ML methods by replacing time-consuming manual knowledge extraction with easier tasks. For example, learning the fitness values for disease descriptions of the ONE from data was shown useful: The knowledge learnt from data produced a knowledge base that performed better with real world cases with confounding values [Viikki and Juhola, 2001]. Methods for weight definition based on data were seen beneficial to implement in the future, not forgetting the need for human experts in the knowledge acquisition process [Viikki and Juhola, 2001]. An equal attribute weighting can, for example, lower the classification accuracy due to noisy, redundant or irrelevant attributes taken into account during the classification [Lee *et al.*, 2007].

This study continues the refinement of the ONE and its knowledge base started in [Viikki and Juhola, 2001; Viikki, 2002]. In the dissertation, machine learning

methods are applied to the knowledge discovery in the otoneurological domain. The aim of the study is to improve the classification accuracy of the decision support system in real world situations with patients having also confounding symptoms. The phases of the dissertation can be divided into three parts: fitness value formation for attribute values, attribute weighting and classification task redefinition (Figure 1).

In the first phase, it is concentrated on the knowledge update of the ONE with the domain experts [I] and on the knowledge discovery method that forms the fitness values for the values of the attributes [I; II]. The knowledge of the ONE needed update due to the update process of data collection questionnaire: New questions were added into the questionnaire and changes were made to answer alternatives of categorical questions and, therefore, the otoneurological paper questionnaire and the decision support system needed to be harmonized and the knowledge base of the ONE updated, if necessary. The effect of the machine learnt fitness values on classification are examined and classification results are compared to the knowledge set by the domain experts and their combinations. In addition, the classification performance of the inference method of the ONE is compared to other classifica-

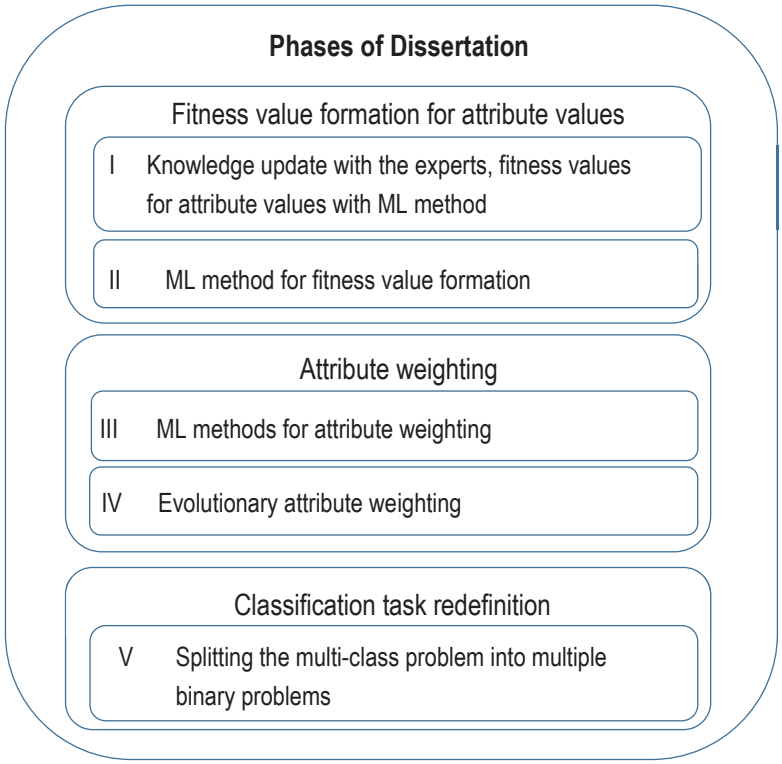


Figure 1. The phases of the dissertation

tion methods (to *k*-nearest neighbour method (*k*-NN) [Cover and Hart, 1967] and Naïve Bayes (NB) [Mitchell, 1997]) [II]. An expanded otoneurological data set and a totally new otoneurological data set are utilized in the study.

The second phase concentrates on the attribute weighting [III; IV]. Different machine learning methods (Scatter method [Juhola and Siermala, 2012] and instance-based learning algorithms IB4 [Aha, 1992] and IB1w (combination of IB1 [Aha *et al.*, 1991] and IB4 methods) are applied in the attribute weighting. These machine learnt attribute weights in addition to the weights defined by the domain experts and equal weighting (all weights set to 1) are tested with different classification methods (with a nearest pattern method of the ONE and an attribute weighted *k*-NN with One-vs-All other (OVA) [Rifkin and Klautau, 2004] classifiers (*wk*-NN OVA)) [III]. An evolutionary approach (genetic algorithm (GA) [Michalewicz, 1992]) is examined in the attribute weighting [IV]. The machine learnt weight sets formed in Publication III are utilized as a starting point with the GA. During the GA runs, the populations are evaluated with the nearest pattern method of the ONE, with the *wk*-NN OVA and with an attribute weighted *k*-NN using neighbour's class-based attribute weighting (*cwk*-NN).

In the third phase, the effect of the classification task redefinition is examined [V]. The multi-class classification task is separated into several binary classification tasks with OVA and One-vs-One (OVO) [Fürnkranz, 2001] classifiers. In this study, the classification is examined without the attribute weighting with the basic *k*-NN and support vector machines (SVM) [Cortes and Vapnik, 1995].

This thesis consists of the present introductory part and five original publications. The remainder of the introductory part of the dissertation is divided as follows. Chapter 2 gives a brief overview of the otoneurological domain and difficulties confronted there in discrimination of the otoneurological diseases. In addition, it presents shortly the otoneurological decision support system ONE. In the end of the Chapter 2, the otoneurological data utilized in the research is described in detail. Chapter 3 presents shortly the machine learning methods and the result evaluation measures utilized in Publications I–V. Machine learning methods are divided into the classification and knowledge discovery methods depending on their role in the research. Overviews and results from the individual publications are introduced in Chapter 4. Discussion and conclusions of the dissertation are presented in Chapter 5.

2 OTONEUROLOGICAL DOMAIN

2.1 Otoneurological Diseases

Vertigo, dizziness and balance disorders are amongst the most common reasons for visiting a physician [Rammazzo *et al.* 2016], even the single most common complaint among patients older than 75 years [Chawla and Olshaker, 2006]. Vertigo affects approximately 20–30% of the general population at any age [Dong *et al.*, 2014] whereas dizziness and balance disorders affect up to 30–40% of the population by 60 years of age [Rammazzo *et al.* 2016]. Dizziness and balance disorders can lead to falls and related fractures causing also other complications and loss of function [Chawla and Olshaker, 2006; Exarchos *et al.*, 2016]. Vertigo, dizziness and balance disorders can be symptoms of many different diseases [Kentala, 1996b]. The characteristics of vertigo usually vary depending on the disease [Baloh, 1995]. The reason for vertigo can be, for example, the vestibular organ, migraine, tumour, infection, head or neck injury, syphilis, medication or chronic alcoholism [Kentala, 1996b; Chawla and Olshaker, 2006]. The causes of vertigo involve hundreds of diseases, which aetiology is associated with otology, neurology and general medicine [Dong *et al.*, 2014]. Thus, the evaluation of the vertiginous patient can be overwhelming for any physician [Chawla and Olshaker, 2006] and, therefore, systems to support and aid diagnosis of vertiginous patients are considered beneficial by the physicians [Aalto, 2005]. Medical history is essential in evaluating the patient's vertigo and in the assessment of the vertiginous patient [Chawla and Olshaker, 2006]. Appropriate clinical examinations based on a patient and symptoms specifically are in a key role with a systematic history taking in forming the diagnosis [Chawla and Olshaker, 2006; Exarchos *et al.*, 2016]. When the symptoms are indistinguishable, clinical tests are needed to confirm the disease [Kentala, 1996b]. For example, acoustic neurinoma can be confirmed with computerized tomography or magnetic resonance imaging [Kentala and Pyykkö, 2000].

In the dissertation, classification and support for the diagnosis of nine otoneurological diseases are concentrated on: acoustic neurinoma (ANE, tumour), benign positional vertigo (BPV), Menière's disease (MEN), sudden deafness (SUD), traumatic vertigo (TRA), vestibular neuritis (VNE), benign recurrent vertigo (BRV),

vestibulopatia (VES) and central lesion (CL). Characteristics of the six first mentioned diseases are presented in [Kentala, 1996a, Kentala, 1996b]. The data of vertiginous patients were collected at the Department of Otorhinolaryngology at Helsinki University Central Hospital, Finland (HUCH) and at the Department of Otolaryngology at Tampere University Hospital, Finland (TAUH). Most of these vertiginous patients were not common cases, rather, they offered diagnostic difficulties for the referring physicians and, therefore, were remitted to more thorough investigations into the departments of otorhinolaryngology [Kentala *et al.*, 1998]. Apparent cases of vestibular neuritis and benign positional vertigo were possible to diagnose already by the general practitioners [Kentala, 1996b] and, thus, seldom occur in the current collected data. The difficulty of the distinguishing the diseases can be seen in the research of Kentala [Kentala, 1996a]: 1167 patients filled out an otoneurological questionnaire but definite diagnosis was possible to give only for 872 patients. Also in [Kentala *et al.*, 1998], ten patients from the original 33 patients were excluded from the research because even the experienced otoneurological experts could not confirm their diagnoses.

Distinguishing different otoneurological diseases causing vertigo from each other can be challenging because there can occur similar kind of and overlapping symptoms with different diseases and, with some diseases, the symptoms can vary over time making recognition difficult [Kentala, 1996b, Havia, 2004]. Some diseases can be said to simulate each other. For example, the main symptoms of acoustic neurinoma are hearing loss and tinnitus but about half of the acoustic neurinoma patients experience also vertigo [Kentala and Pyykkö, 2001]. In the research of Kentala and Pyykkö [Kentala and Pyykkö, 2000], over third of the acoustic neurinoma patients had the full triad of vertigo, hearing loss and tinnitus that are characteristic for the Menière's disease. From these acoustic neurinoma patients, 14% reported their vertigo to mimic the vertigo encountered in Menière's disease. The vertigo of acoustic neurinoma patient can also be similar to benign positional vertigo. In addition, the hearing loss of acoustic neurinoma patients can mimic sudden deafness or it can fluctuate like in Menière's disease. Sudden deafness and Menière's disease can have similar kind of symptoms in the beginning and only a follow-up will reveal, which disease is in question [Kentala *et al.*, 1998]. When the Menière's disease progresses, almost half of the patients develop bilateral auditory symptoms and vertigo attacks occur more frequently and are more severe than in the beginning of the disease [Chawla and Olshaker, 2006]. Also, traumatic vertigo patients can have symptoms typical of Menière's disease [Havia, 2004] or they can mimic symptoms of benign positional vertigo [Kentala, 1996a]. Moreover, the actual

disease can change over the course of time [Kentala *et al.*, 1998]. For example, after 8.5 years follow-up of benign recurrent vertigo patients, 14.0% of cases had evolved BRV to typical Menière's disease, 8.0% had benign positional vertigo, 10.0% has still active benign recurrent vertigo and 62.0% had no vertigo anymore [Rutka and Barber, 1986]. Also, 6.0% of vestibular neuritis patients have been reported to develop benign positional vertigo during the recovery period [Kentala, 1996b] and up to 22.0% of acoustic neurinoma cases have had sudden deafness [Kentala, 1996a]. In addition, patients can have symptoms and signs that are not related to the current disease, so-called confounding symptoms, that makes the diagnosis even more demanding [Kentala, 1996b]. Confounding symptoms can be, for example, noise-induced or age-related hearing loss, medication or chronic disease causing additional symptoms.

The most important questions in discriminating between the six most common otoneurological diseases were the occurrence (frequency) and duration of the vertigo attacks, the duration of hearing loss, the duration of vertigo and the occurrence of head injury [Kentala, 1996a]. These attributes were also within the most important attributes defined by the decision tree [Viikki, 2002]. Baloh considered the most important questions for evaluating a dizzy patient the duration and occurrence of vertigo, the symptoms aggravated by head movement and the auditory or neurologic symptoms associated with it [Baloh, 1995]. The value distributions of attributes 'ATT_OFTEN: occurrence of stronger vertigo attacks' and 'ATT_LAST: duration of vertigo attacks' by the disease classes within the HUCH data containing nine disease classes are presented in Figures 2 and 3. Even though these two attributes are considered the most important questions in distinguishing the vertigo diseases, it can be seen that there are similar kind of distributions with different diseases, which makes it difficult to separate them from each other.

In order to understand the difficulty of the differentiation of otoneurological diseases better, a principal component analysis (PCA) was made to the otoneurological data collected at HUCH containing all nine disease classes during the dissertation study. The PCA was made with SPSS Statistics software. The PCA reveals if there exists components that are useful for representing the whole data [Duda *et al.*, 2001]. The PCA generates a new set of attributes, principal components that are linear combinations of the original attributes. Thus, it is possible to reduce the dimensionality of the data set [Duda *et al.*, 2001]. The otoneurological data is high-dimensional with 94 attributes. The total variance explained by the first principal component was 12.5%, by the second principal component 9.2% and by the third principal component 4.5%. Thus, the two main principal components

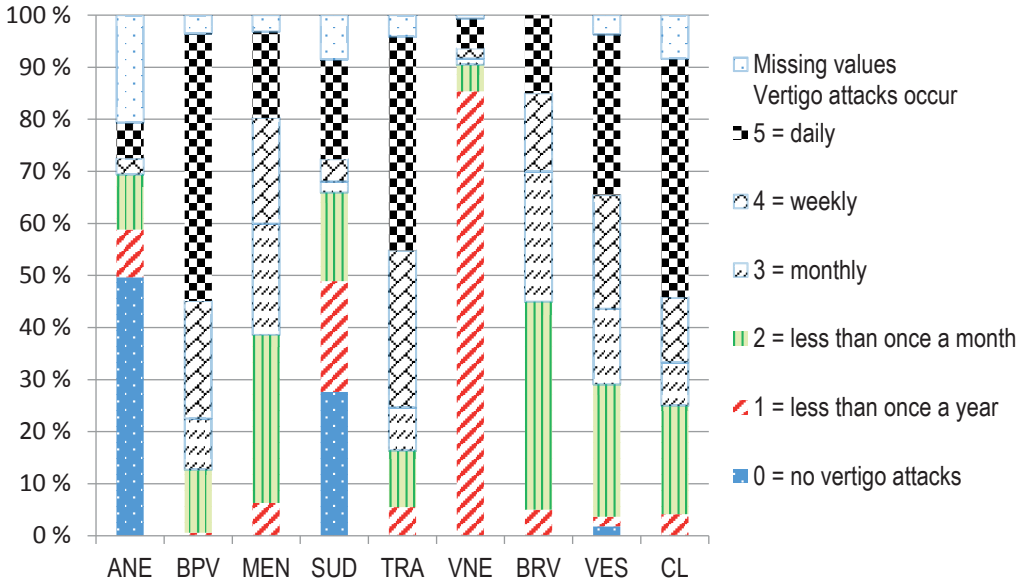


Figure 2. The value distribution of attribute 'ATT_OFTEN: occurrence of stronger vertigo attacks' within augmented HUCH data containing nine disease classes

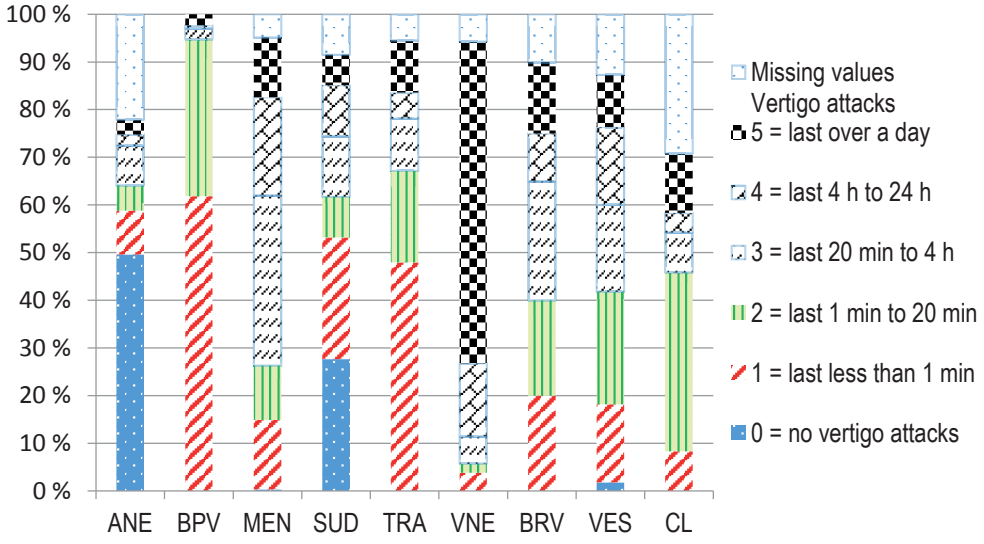


Figure 3. The value distribution of attribute 'ATT_LAST: duration of vertigo attacks' within augmented HUCH data containing nine disease classes

explained only 21.7% of the total variance. The augmented HUCH data projected onto the first and the second main principal components are visualized in Figure 4.

As it can be seen from the projection, most of the disease classes are extensively overlapping with each other. Most of the acoustic neurinoma cases are obviously separate from the other disease classes except sudden deafness. This forms the first cluster. Some of the sudden deafness cases are mixing with the Menière's disease cases that form the basis of the next cluster. Almost a half of the traumatic vertigo, vestibulopatia, benign positional vertigo and central lesion cases are overlapping with Menière's disease. Also, some of the vestibular neuritis cases can be found from this cluster. The third overlapping cluster is mostly separate from Menière's disease, acoustic neurinoma and sudden deafness: Vestibular neuritis, benign positional vertigo, benign recurrent vertigo, traumatic vertigo, vestibulopatia and central lesion cases are overlapping with each other. The strong overlap with the disease classes makes the separation of them challenging.

In order to understand the result of the PCA better, a closer look of the principal components was taken. The first main principal component can be called a hearing disorder component because it concentrated on hearing loss and tinnitus in addition to audiometry measurements. The highest principal component coefficients within the first principal component were achieved with the attributes 'LAT_KA: bi- or unilateral hearing loss', 'NONLAT_KA: normal hearing' (negative influence), 'SYM_HEARLOSS: do you have hearing loss', 'AGE_HL_SYM: age of hearing loss', 'HL_TYPE: type of hearing loss' and 'HL_SIDE: side of hearing loss'. The second main principal component concentrated on vertigo, vertigo attacks and headache during them. The highest principal component coefficients were achieved with the attributes 'VERTIGO: true vertigo containing feeling of rotation or floating', 'SYM_VERT: do you have vertigo', 'BILATERAL: bilateral hearing loss/ tinnitus/ hyperacusis/ pressure feeling in the ear', 'UNILATERAL: unilateral hearing loss/ tinnitus/ hyperacusis/pressure feeling in the ear' (negative influence), 'LIGHTHEAD: do you have feeling of unreality' and 'ATT_OFTEN: occurrence of stronger vertigo attacks'. The third main principal component concentrated on ear and head trauma. The highest principal component coefficients were achieved with the attributes 'TRAUMA: serious trauma of the head', 'HEAD_TRAUMA: direct trauma to the head or neck associated with the beginning of the vertigo symptoms (occurred within 6 months)', 'INJURY: head or ear trauma, noise injury' and 'CONCUSSION: brain concussion with unconsciousness lasting less than 2 hours'.

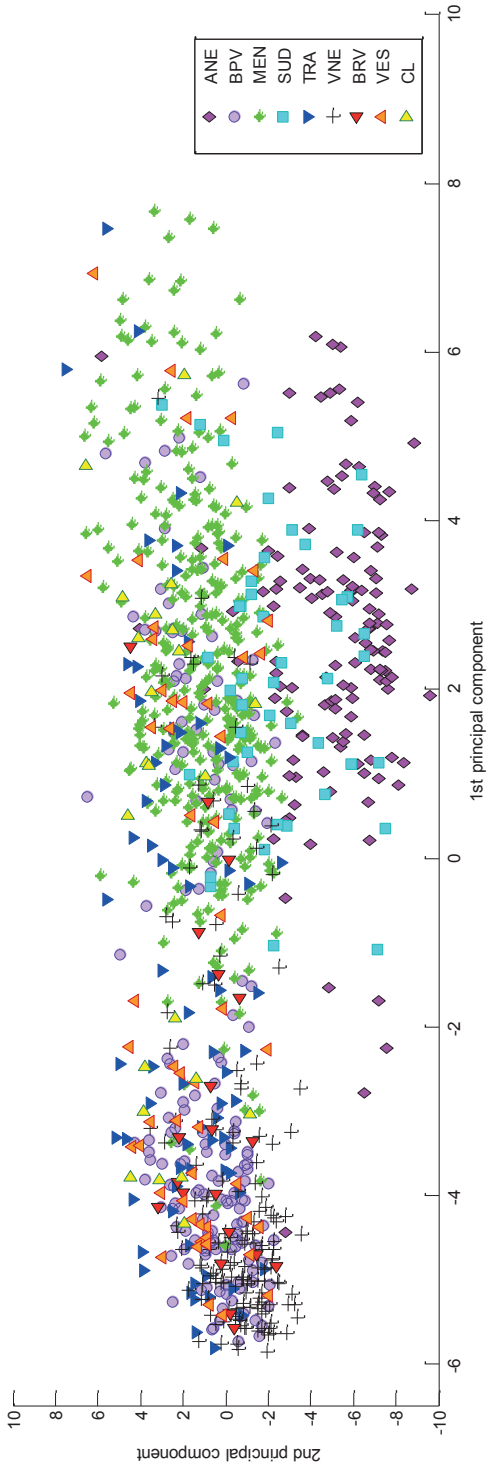


Figure 4. Nine disease classes of augmented HUCH data projected onto two main principal components defined by the principal component analysis. The first principal component concentrated on hearing disorders and the second on vertigo disorders.

2.2 Otoneurological Decision Support System ONE

As previous chapter presented, otoneurological diseases can be difficult to separate from each other because of their overlapping and similar kinds of symptoms. Therefore, the development of the otoneurological decision support system ONE [Kentala, 1996b; Auramo, 1999] was started in the 1990s. The ONE was targeted to general practitioners to assist the diagnosis of vertigo, to specialists for collecting information on vertiginous patients and to medical students to work as an educational tool [Kentala, 1996b]. Inference of the ONE is based on patient history, occurring symptoms and results of the clinical tests [Kentala *et al.*, 1996]. The knowledge utilized in decision making of the ONE is presented in its knowledge base by disease-wise descriptions in the form of weights for the attributes and fitness values for the values of the attributes [Auramo *et al.*, 1993]. The weights describe the significance of the attributes for the disease. This kind of knowledge representation model for the ONE was selected due to its intelligibility: It was wanted that the knowledge representation would be in a form that could be easily understood by humans and the knowledge should be easily fixed or tuned in the future, for example, with machine learning methods [Auramo, 1999]. The original knowledge base of the ONE was constructed with the help of experienced otoneurologists and on the data obtained from the literature [Kentala *et al.*, 1998]. It included both central and peripheral vertigo diseases, at total 18 disease descriptions.

The first version of the ONE was made in C++-language. Already Auramo acknowledged the need to transfer the ONE to Windows environment due to MS-DOS memory limitations [Auramo, 1999]. The upgrade process of the ONE was started in the 2000s by Tapani [Tapani, 2008] by transferring the ONE to the Java environment. Tapani further developed the graphical user interface of the ONE, added a batch processing possibility to run in the Windows command prompt with parameters and programmed the *k*-nearest neighbour method [Mitchell, 1997] into the ONE. The upgrade and refinement process of the ONE was continued by the author with addition of MySQL database functionality into the system and update of the questionnaire [Varpa, 2005, I]. The more detailed description of the upgrade and refinement process and the state of the ONE after the process is given in Chapter 4.1.

The ONE consists of separate components: a graphical user interface, a query base, an answer database, a knowledge base and an inference engine (Figure 5). The graphical user interface is created on the basis of the query base. The query base contains information about the questions to be shown at the same view in the user

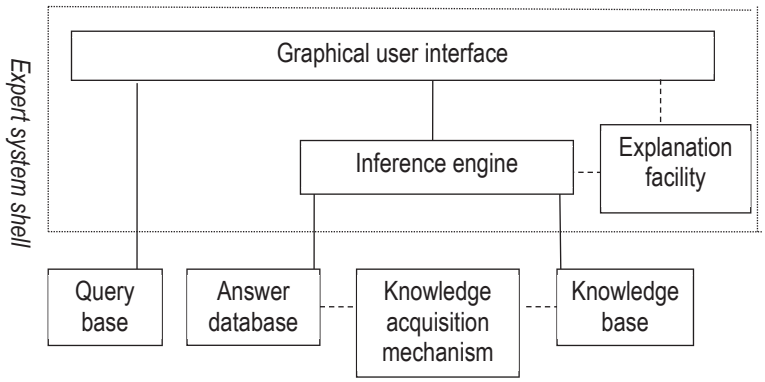


Figure 5. Main components of the otoneurological decision support system ONE [I]

interface, the type of questions and the possible answer alternatives for the questions. The views in the query base are divided into symptoms, medical history, clinical findings and life quality. In the symptoms part, questions concern, for example, vertigo, gait difficulties, hearing loss, tinnitus and hyperacusis. The medical history contains questions about drug usage, head and ear injury, ear operations and other diseases. In the clinical findings, the results of different otoneurologic, audiologic and imaging tests are given. The answer database stores the data about the patient in the MySQL database. The tables of the database can be created on the basis of the query base (one view corresponds to one table). The knowledge base contains descriptions (patterns) of deducible classes in the form of attribute weight values and fitness values for the values of the attributes (Figure 6). The weight value of the attribute expresses the significance of the attribute for the class and the fitness values show what attribute values fit for the class. The weight values vary from zero to a chosen maximum, where zero means that the attribute does not concern the class at

(a) <attribute name> <weight value> <attribute type>
<minimum value> <maximum value>
<value 1> < fitness value 1>

...

<value *n*> < fitness value *n*>
END
.

(b) ATT_OFTEN 4 V
0.0 5.0
0.0 0.0
1.0 1.12
2.0 23.60
3.0 19.10
4.0 43.82
5.0 100.0
END
.

Figure 6. (a) General form of an attribute pattern in the knowledge base of the ONE and (b) an example attribute description 'ATT_OFTEN: frequency of vertigo attacks' with benign positional vertigo [III]

all. The greater the weight value, the more important the attribute is for the class. Fitness values can have values between zero and 100.

The fitness value zero means that the attribute value does not fit the class whereas the fitness value 100 shows that the value is the most typical for the class. Each attribute in the knowledge base refers to a sign, a symptom or a measurement data from a clinical test. In the otoneurological domain, the descriptions in the knowledge base can be considered as a profile of a disease class, which describes symptoms and signs related to the disease. It matches partly the clinical picture of the disease but, in the description, it is possible to take into account also unusual symptoms and signs that a patient may have. The inference engine uses the weights and the fitness values given in the knowledge base in its inference. The inference engine resembles the nearest neighbour method [Mitchell, 1997] but, instead of the nearest case, it searches for the nearest pattern from the knowledge base [Auramo and Juhola, 1996]. Therefore, the inference engine of the ONE is referred to the nearest pattern method within the dissertation and in the publications.

In addition, the ONE contains an explanation facility that informs the user if some answers do or do not fit for the class or if some crucial information is missing in order to make the diagnosis. Explanations are based on the information given in the knowledge base, for attribute type. Some attributes are defined as necessary (V) and others supporting (T). Originally, necessary attributes were definitely required in the case of the disease: If a patient did not have a symptom described as necessary, the disease was inferred to be impossible [Auramo *et al.*, 1993]. At that time, the disease patterns in the knowledge base of the ONE were considered as clinical pictures of the diaseses and, therefore, only the symptoms related on the disease at hand were taken into account. Later, the patterns were expanded to take into account also symptoms relating to other diseases and confounding symptoms. The use of necessary attribute values in the original way could reject the correct diagnosis due to the unfitting necessary attribute values [Viikki, 2002]. This weakened especially the classification of the BPV and VNE. Therefore, necessary attribute values attached to certain disease descriptions in the knowledge base of the ONE were not applied to reject disease in the inference of the ONE anymore in [Viikki and Juhola, 2001] and later in the studies in Publications I–IV. However, these necessary attribute values are still utilized in the explanation facility of the ONE to show the user how the ONE ended up for its classification results, which symptoms do or do not fit for the class.

The graphical user interface, the inference engine and the explanation facility of the ONE form an expert system shell that is possible to take into use in different

domains. The query base and the knowledge base need to be adjusted into the new domain. The help of domain experts is needed for data collection and domain knowledge. After the collection of domain data, it is possible to utilize machine learning methods presented in the dissertation in creation of the domain knowledge for the knowledge base.

2.3 Otoneurological Data

Gathering of HUCH data set was started during the development of ONE in the beginning of the 1990s [Kentala *et al.*, 1995] and it continued over a decade [Viikki, 2002]. The first collected data set (564 cases) included only the cases that had no confounding symptoms from six major patient groups: acoustic neurinoma (aka vestibular schwannoma), benign positional vertigo, Menière's disease, sudden deafness, traumatic vertigo and vestibular neuritis [Kentala, 1996b]. This data set was utilized to study more, which symptoms and findings were characteristic for these six disease classes. Four data sets containing also cases with confounding values were added into the HUCH data during the years, thus, forming the current augmented data set of 1030 cases. The more detailed description of the distributions of the disease classes within the first four otoneurological data sets can be found from [Viikki, 2002]. Before collecting otoneurological data in the TAUH, a refinement of the otoneurological questionnaire was made with the domain experts [I]. The TAUH data set was gathered during the years 2004 and 2005.

Table 1. The frequency distributions of the disease classes in the HUCH and TAUH data

			HUCH data		TAUH data	
	Disease name	**Abbreviation**	***n***	**%**	***n***	**%**
1	Acoustic Neurinoma	ANE	131	12.7		
2	Benign Positional Vertigo	BPV	173	16.8	80	31.6
3	Menière's Disease	MEN	350	34.0	128	50.6
4	Sudden Deafness	SUD	47	4.6		
5	Traumatic Vertigo	TRA	73	7.1		
6	Vestibular Neuritis	VNE	157	15.2	20	7.9
7	Benign Recurrent Vertigo	BRV	20	1.9		
8	Vestibulopatia	VES	55	5.3	25	9.9
9	Central Lesion	CL	24	2.3		
			1030	100.0	253	100.0

At total, the utilized augmented HUCH data set contained information and diagnoses on 1030 patients from nine otoneurological disease classes and the TAUH data set on 253 patients from four otoneurological disease classes (Table 1). Each case was a patient who had informed to have vertigo or gait difficulties and was diagnosed to have one of the listed otoneurological diseases. Patients filled out an otoneurological questionnaire with 105 questions concerning their symptoms and medical history. The diagnoses of the patients were confirmed by the experienced specialists [Kentala, 1996b] who could use patient records in addition to the otoneurological questionnaire in making a diagnosis of a patient.

The frequency distributions of the disease classes were imbalanced with both the data sets. Majority of the cases belonged to the class Menière's disease: Over third of the cases in the HUCH data set (34.0%, 350 cases) and over half of thecases in the TAUH data set (50.6%, 128 cases) belonged to this class. One reason for the large number of Menière's disease cases in the TAUH data was that there was a co-operation project on peer-support going on with the Finnish Menière Federation at that time. The smallest disease classes used in the studies contained only 20 cases. Due to the small number of disease classes and the highly imbalanced class distribution, TAUH data set was used only in Publication II.

In Publications II, III and V, the HUCH data set was used with all 1030 cases from nine disease classes within the classification runs using only the machine learnt knowledge. During the classification runs in Publications I, II, III and IV where the experts' knowledge was compared with the machine learnt knowledge, the HUCH data set with 951 cases from seven disease classes were used. The experts could define the attribute weights and fitness values only for seven disease classes: Two classes (vestibulopatia and central lesion) were found to be too complex to be described with the attribute weights and fitness values at our disposal.

2.3.1 Attributes

The otoneurologic questionnaire contained 105 questions about the patient's symptoms and medical history, for example, about vertigo, gait difficulties, hearing loss, tinnitus, hyperacusis, drug usage, head and ear injuries and other diseases. The refined questionnaire [I] can be found from the Appendix I. It included also questions about family history of vertigo and hearing loss and fifteen questions about the health-related quality of life (15D HRQoL) questionnaire [Sintonen, 2001], but, these questions were only answered by the TAUH patients. Therefore, we did not

include these new attributes in the research. In addition to the otoneurological questionnaire, the decision support system contained 72 questions concerning otoneurologic, audiologic and imaging tests, for example, audiometry frequencies and posturography, and attributes derived from the answers from the questionnaire and the clinical tests. The clinical tests were not performed on each patient and, therefore, values of the attributes were missing in several test results. Especially in the HUCH data, no clinical examinations were made solely in order to collect data for the study, only the necessary examinations for the patients were ordered by the physician in charge [Kentala, 1996b]. In total, there were 177 questions related to a HUCH patient and 192 questions related to a TAUH patient. In earlier otoneurological studies, 38 attributes were defined central by the domain experts [Kentala *et al.*, 1999].

Attributes with low frequencies of available values were discarded from the data set. In Publications I and II, any attribute was discarded if it had over 30% values missing and in Publications III, IV and V any attribute was discarded if it had over 35% values missing. There was one exception for this: The attribute 'HL_TYPE: type of hearing loss' was kept in the data set even it had 52.8% of its values missing in total. This attribute is essential in the recognition of sudden deafness [Kentala, 1996a] and, therefore, it could not be discarded from the data set. One reason for the high rate of missing attribute values for the attribute hearing loss type can be that this question was not asked in the original paper questionnaire, only in the decision support system. With other diseases, hearing loss usually starts during several months but with sudden deafness the hearing loss starts suddenly. The value distributions of hearing loss type within disease classes are presented in Figure 7. The values of attribute HL_TYPE were missing mainly from the patients suffering from other diseases than sudden deafness and, thus, the values were not missing totally at random. After discarding the attributes with low frequencies of available values, 89 attributes were used in Publications I and II and 94 attributes in Publications III, IV and V. From these 94 attributes, almost half (48.9%, 46 attributes) had less than 5% missing values and 77.7% (73 attributes) had less than 10% missing values.

In Publications I and II, the continuous attributes were discretized into equal-width intervals and, thus, all the 89 attributes were of the ordinal type. Discretization of continuous attributes was done in order to have similar kind of attribute handling with all machine learning methods used in Publication II. In Publications III, IV and V, 17 attributes were quantitative (integer or real value) and 77 attributes were qualitative (from which 54 were binary (yes/no), 20 were ordinal and 3 were nominal).

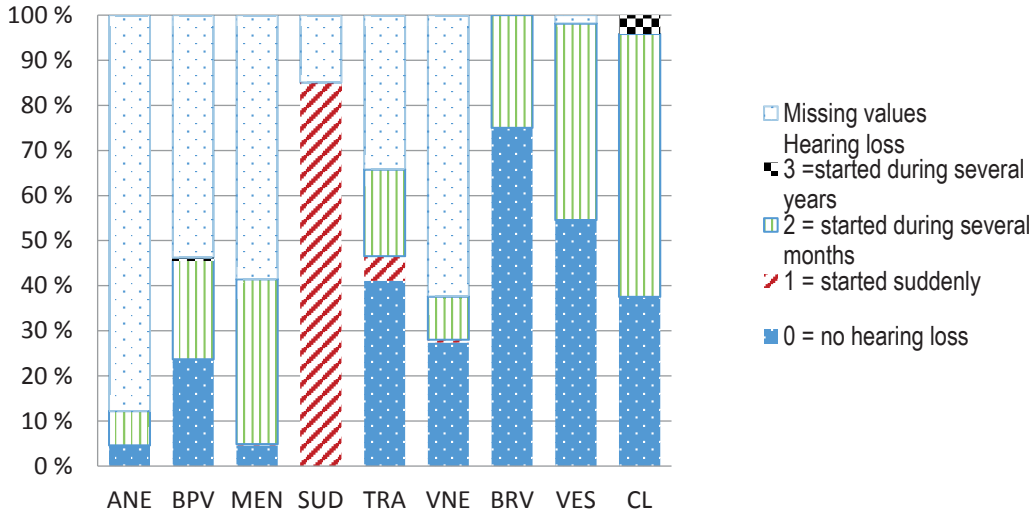


Figure 7. The value distribution of attribute 'HL_TYPE: type of hearing loss' within HUCH data containing nine disease classes

The augmented HUCH data set with missing attribute values was utilized in the fitness value calculations and in the classification runs in Publications I, II, III and IV. Due to the Scatter method [Siermala *et al.*, 2007; Juhola and Siermala, 2012] and the support vector machines [Cortes and Vapnik, 1995], the data was necessary to impute (replace missing values) because these methods require complete input data in order to work properly. There were only 22 complete cases (2.1%) in the augmented HUCH data set when 94 attributes were considered and, thus, the training set would have been too small without imputation. The small percentage of missing attribute values (9.8%) allowed the use of imputation. The imputation was done class-wise on the basis of the whole data set prior to data division into training and testing sets. The missing values of the attributes were imputed (substituted) with the class modes of the qualitative attributes (binary and nominal) and with the class medians of the quantitative attributes (ordinal, interval scale and ratio scale). Laurikkala *et al.* [2000] showed that these simple imputation methods were adequate enough for otoneurological data. The imputed data was used with the attribute weighting methods in Publication III and with the classification methods in Publication V.

3 MACHINE LEARNING

In this chapter, the machine learning methods utilized in the dissertation are presented shortly. Before that, some key concepts of the dissertation from the machine learning area are defined. The utilized ML methods are divided into classification methods and knowledge discovery methods depending on their role in the research. In the end of the chapter, measures used for result evaluation in the research are presented.

3.1 Definitions

Machine learning (ML) can be broadly defined as "any computer program that improves its performance at some task through experience" [Mitchell, 1997]. The objective in ML is to develop computational methods that would implement different ways of learning to induce knowledge from data [Kubat *et al.*, 1998]. In other words, ML is about algorithms that infer structure from data and ways to validate that structure; it is a technology for mining knowledge from data [Witten *et al.*, 2011].

Knowledge discovery (KD) can be defined as "a process of discovering meaningful patterns in data automatically (or semi-automatically), which help to explain something in data" [Witten *et al.*, 2001]. This can be called also data mining. The patterns are considered meaningful when they lead to some advantage, usually an economic one [Witten *et al.*, 2001]. Meaningful pattern can contain, for example, important regularities occurring in data, it can show the characteristics of the class or other useful information about the domain and relationships occurring there. In the dissertation, ML methods utilized in forming the fitness values for the attribute values and the weight sets for the attributes from the otoneurological data are referred to as knowledge discovery methods.

The classification methods can be defined as ML methods that are utilized in classifying example cases into one of a discrete set of possible categories when each case is described with a vector of attribute values [Mitchell, 1997]. Classification can

be described as "predicting a label from the predefined set of classes for the unknown objects" [Joutsijoki, 2012].

Attribute weighting is needed to grade the relevancy and usefulness of the attributes: In worst case, noisy, redundant and/or irrelevant attributes may reduce the classification accuracy when treating all attributes as equally important (with unweighted attributes) [Lee *et al.*, 2007].

3.2 Classification Methods

3.2.1 Nearest Pattern Method of the ONE

The main classification method used within the research was the nearest pattern method of the decision support system ONE, which classification performance was studied with different fitness values of the attribute values [I; II] and attribute weight sets [III; IV]. The nearest pattern method of the ONE [Auramo and Juhola, 1996] searches for the best fitting pattern (class) from its knowledge base. It calculates scores for the patterns from the attribute weight values and fitness values of the attribute values. The score $S(c)$ for a class c is calculated in the following way:

$$S(c) = \frac{\sum_{a=1}^{A(c)} x(a)w(c,a)fv(c,a,v)}{\sum_{a=1}^{A(c)} x(a)w(c,a)}, \quad (1)$$

where $A(c)$ is the number of the attributes associated with the class c,
$x(a)$ is 1 if the value of the attribute a is known and otherwise 0,
$w(c,a)$ is the weight of the attribute a for the class c and
$fv(c,a,v)$ is the fitness value for the value v of the attribute a for the class c.

In the case of quantitative attributes, the fitness values are interpolated by using the attribute values in the knowledge base as interpolation points. The fitness values are altered to the range of 0 to 1 during the inference process. The class pattern having the highest score is the best diagnosis suggestion. In addition to the scores calculated from the known information, minimum and maximum scores for the patterns are calculated using the lowest and the highest fitness values of the attribute values with the attributes having missing values. With the minimum and maximum scores, the ONE handles uncertainty caused by the missing attribute values. The closer the minimum and maximum scores are to each other, the more reliable the inference is. The diagnosis suggestions of the ONE are ordered primarily by the score and

secondarily by the difference of the minimum and maximum scores. If the patterns have the same score but one pattern has a smaller difference between the minimum and maximum scores than the others, the pattern having the smallest difference is placed as a higher diagnosis suggestion. If the patterns have the same score and the minimum and maximum score difference, their order is selected randomly. However, this situation occurs rarely.

In Publication I, the classification runs with the ONE using the knowledge bases containing machine learnt fitness values were processed manually in the Windows command prompt where it was possible to make batch processing with more than one patient case. Data sets and knowledge combinations utilized within the classification were input manually to the ONE and started separately in the command prompt. This was the reason why the 10-fold cross-validation (CV) was repeated only once. Afterwards, a batch file (script file) was created by the author to help the classification runs of the ONE (both the nearest pattern method and the *k*-NN implemented into the ONE system) in order to avoid input errors and to speed up the runs by allowing the continuous processing of CV runs. Thus, in Publication II it was possible to repeat the 10-fold CV runs three times with the ONE using machine learnt knowledge bases (KB2–KB5) and with the *k*-nearest neighbour method. Later, the nearest pattern method of the ONE, the basic *k*-NN and the attribute weighted *k*-nearest neighbour with OVA classifier (*w*k-NN OVA) were transferred to the Matlab by the author. The classification runs of the ONE in Publications III and IV were run in the Matlab, thus, allowing to repeat the 10-fold CV 10 times and even 100 times during the GA runs.

3.2.2 *k*-Nearest Neighbour Method

The nearest pattern method resembles the nearest neighbour method and, thus, it was natural to compare it to the *k*-nearest neighbour method (*k*-NN) [Cover, Hart, 1967]. The *k*-NN is simple to compute because it is an instance-based (case-based) learning method [Mitchell, 1997]. The *k*-NN explains its inference by showing the *k* nearest cases. This is analogous to the way how domain experts make diagnosis on the basis of previously known similar cases [Kononenko *et al.*, 1998]. The *k*-NN searches for the *k* most similar (nearest) cases from the training set and classifies a new case into the most frequent (majority) class of these *k* training cases (Figure 8). If there is more than one majority class, the predicted class is selected randomly from the majority classes.

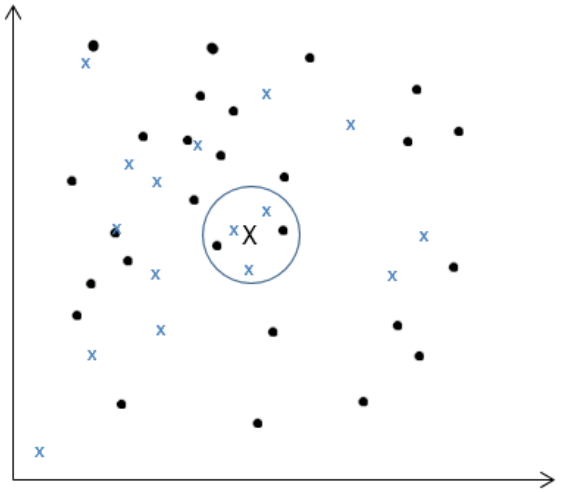


Figure 8. In the *k*-nearest neighbour method, the *k* most similar cases from the training set are searched for and the new case *X* is classified into the most frequent class of these *k* training cases.

The most similar (the nearest) cases are solved with a distance measure in the *k*-NN. In Publication II, the Value Difference Metric (VDM) [Wilson and Martinez, 1997] was used as a distance measure whereas, otherwise, the Heterogeneous Value Difference Metric (HVDM) [Wilson and Martinez, 1997] was used. The VDM can handle only qualitative attributes whereas the HVDM handles both qualitative and quantitative attributes in the data set. The HVDM is defined as

$$HVDM\ (x, y) = \sqrt{\sum_{a=1}^{n} d_a(x_a, y_a)^2}, \tag{2}$$

where n is the number of the attributes and
$d_a(x_a,y_a)$ is the distance between the values x_a and y_a for the attribute a.

The distance function $d_a(x_a,y_a)$ is defined as

$$d_a(x_a, y_a) = \begin{cases} 1, & if\ x\ or\ y\ is\ unknown \\ normalized\_vdm_a(x_a, y_a), & if\ a\ is\ qualitative \\ normalized\_diff_a(x_a, y_a), & otherwise \end{cases}. \tag{3}$$

Because HVDM computes distances to the qualitative and other attributes with different measurement ranges, it is necessary to scale their results into approximately the same range in order to give each attribute a similar influence on the overall distance (normalize results). The normalized distance to the qualitative attribute is calculated as

$$normalized\_vdm_a(x_a, y_a) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2}, \quad (4)$$

where C is the number of output classes in the problem domain,
$N_{a,x(y),c}$ is the number of cases that have the value x (or the value y) for the attribute a and the output class c and
$N_{a,x(y)}$ is the number of cases that have the value x (or the value y) for the attribute a

and to the quantitative attribute as

$$normalized\_diff_a(x_a, y_a) = \frac{|x_a - y_a|}{4\sigma_a}, \quad (5)$$

where σ_a is the standard deviation of the numeric values of the attribute a in the training set of the current classifier (Wilson and Martinez, 1997).

The basic *k*-NN with unweighted distance measures were utilized in Publications II and V. The classification runs were made with the basic *k*-NN implemented with Java (the *k*-NN included in the ONE made by Tapani) and, thus, the runs were driven in Windows command prompt with the batch file made by the author. Later, the *k*-NN was transferred to the Matlab by the author.

3.2.3 Naïve Bayes

The Naïve Bayes (NB) classifier [Witten *et al.*, 2011; Mitchell, 1997] was selected to be used in comparison of the ONE in Publication II because the NB classifier is simple, efficient and robust to noisy data [Yang, Webb, 2002]. In addition, its performance has been shown to be comparable to decision tree learning and neural networks in some domains [Mitchell, 1997; Kononenko *et al.*, 1998]. Also, physicians have estimated the knowledge presentation of the NB good; the conditional probabilities interests physicians [Kononenko *et al.*, 1998].

The NB classifier has a probabilistic approach to the classification: A new case is classified into the class with maximal calculated probability. In probability ratio calculations, only occurring attribute values of a training case are taken into account in frequency counting; attributes with missing values are omitted [Witten *et al.*, 2011]. The NB classifier assumes that the attribute values are conditionally independent for the given class.

The Laplace-estimate [Cestnik, 1990] was used for estimation of prior probabilities and the M-estimate for estimation of conditional probabilities [Cestnik,

1990; Mitchell, 1997]. In order to avoid the normal distribution assumption for the continuous attributes of the NB, the continuous attributes were discretized [Witten *et al.*, 2011] into equal-width intervals [Yang and Webb, 2002] in the research in Publication II.

The Naïve Bayes was implemented with the Matlab by Kati Iltanen.

3.2.4 Attribute Weighted *k*-Nearest Neighbour Method with OVA classifiers

In the machine learnt attribute weighting research [III], the distance measure of the *k*-NN was extended to take into account the attribute weights in its distance calculation [Kelly and Davis, 1991]. The attribute weighting was added into the HVDM:

$$weighted\_HVDM\ (x, y) = \sqrt{\sum_{a=1}^{n} w_{c_a} d_a(x_a, y_a)^2}\,, \tag{6}$$

where n is the number of the attributes and
w_{c_a} is the weight of the attribute a in the class c and
$d_a(x_a, y_a)$ is the distance between the values x_a and y_a for the attribute a (described in Equation 3).

The weighted HVDM can handle both qualitative and quantitative attributes.

In order to utilize the same class-wise attribute weight sets in the classification as the nearest pattern method of the ONE, One-vs-All other (OVA, aka 1-vs-All, one-against-all or one-vs-the rest) [Rifkin and Klautau, 2004] binarization was applied with the attribute weighted *k*-NN. In OVA binarization, a multi-class classification task is converted into multiple binary classifiers [Galar *et al.*, 2011]: One classifier is learnt for each deducible class by discriminating that class against the remaining classes (Figure 9). In other words, each OVA classifier is trained to separate a class from all the other classes by marking the cases of this one class as member cases and the cases of the other classes as non-member cases in the training set. The number of classifiers is the number of deducible classes.

The method combining the attribute weighted *k*-NN and the OVA classifiers is called an attribute weighted *k*-nearest neighbour method with OVA classifiers (*wk*-NN OVA). The *wk*-NN OVA searches for the *k* most similar cases of a new case from each classifier separately. Each classifier gives a vote for the new case being a member or non-member of the class based on the majority class of the *k* neighbours. If there is only one classifier suggesting a member of the class, the final class of the new case is assigned from the classifier suggesting the case being a member of the

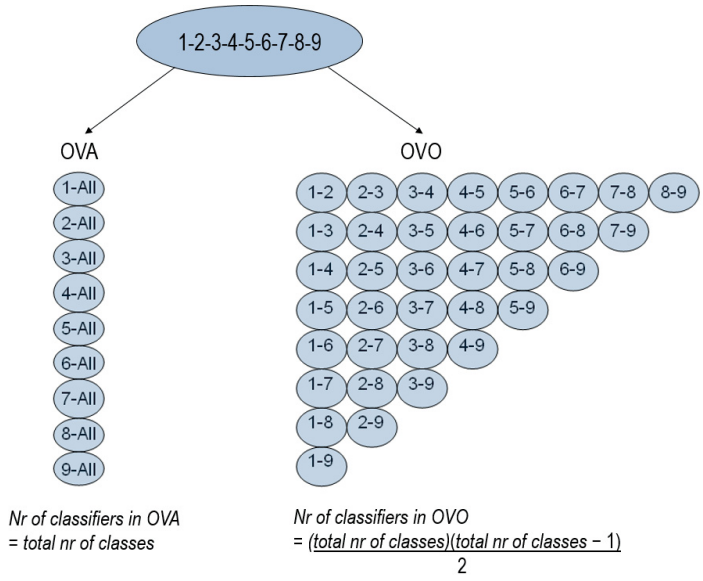


Figure 9. There are two different ways to split a multi-class classifier into multiple binary classifiers: One-vs-All other (OVA) and One-vs-One (OVO) binarization

class (winner-takes-all rule). If there are more than one classifier suggesting it to be a member of the class (a tie situation), the class of the new case is determined by searching the most similar member case from the member voting classifiers (the case with the minimum distance to the new case). However, there can occur also a situation when all the classifiers vote the new case to be the other class, a non-member of all the classifiers. In this case, the class of the new case is determined with the basic 1-NN using the whole training data containing the original disease classes of the cases. In Publication III and IV the attribute weighted 1-NN and in Publication V the unweighted basic 1-NN was used in search for the nearest case.

The attribute weighted *k*-NN with the OVA classifiers was utilized in Publications III and IV. The classification runs were made with the *wk*-NN OVA implemented with Java (with batch files and the *k*-NN included in the ONE) for the Publication III and implemented with the Matlab by the author for the Publication IV.

3.2.5 Attribute Weighted *k*-Nearest Neighbour Method Using Neighbour's Class-Based Attribute Weighting

Since the nearest pattern method of the ONE and the *wk*-NN OVA handled the attribute weight sets separately for each class, it was necessary to take in use a classification method that could handle the class-wise attribute weights at the same time in order to evolve and mutate the whole weight set at a time during genetic algorithm runs. Therefore, for the population evaluation in Publication IV, an attribute weighted *k*-nearest neighbour method using neighbour's class-based attribute weighting (*cwk*-NN) was developed based on a similar kind of class-dependent attribute weighted *k*-NN [Lee *et al.*, 2007].

As with the *wk*-NN OVA, the *cwk*-NN uses the attribute weighted distance measure, the HVDM [Wilson and Martinez, 1997] expanded with the attribute weighting (Equation 6). The difference between the *wk*-NN OVA and the *cwk*-NN is that the *cwk*-NN resembles more the basic *k*-NN: It uses only one classifier instead of multiple classifiers in the classification. The attribute weights to be utilized in the distance calculation between a new case and the training case depend on the class of the training case (*i.e.* the neighbour case). Thus, different sets of attribute weights are used with the training cases belonging to the different classes. There are as many attribute weight sets as there are classes.

The *cwk*-NN was implemented with the Matlab by the author.

3.2.6 Unweighted *k*-Nearest Neighbour Method with OVA and OVO Classifiers

Another approach besides the attribute weighting to improve the classification accuracy might be the classification task redefinition where multi-class classification task is separated into multiple binary classification tasks. With the multiple binary classifiers, it is possible to achieve better understanding about the relationships and differences of the classes and, thus, enhance the understanding of the data and domain at hand [Friedman, 1996; Allwein *et al.*, 2000]. There are two commonly used approaches for splitting a multi-class classification task into multiple binary problems, One-vs-All other (OVA) [Rifkin and Klautau, 2004] and One-vs-One (OVO, aka 1-vs-1, round robin or pair-wise) [Fürnkranz, 2001] binarization. With the OVA classifiers, it is possible to distinguish a separable class from the other classes, if there exist any. Also, the OVO classifiers aid to find out distinguishable classes from the other classes. In Fürnkranz's research [Fürnkranz, 2001], the OVO

classifiers yielded significant improvements in the predictive classification accuracy compared to the OVA classifiers.

In order to examine whether the separation of a multi-class classification task into multiple binary classification tasks affects the classification results, it was necessary to take in use the unweighted versions of the classification methods. Thus, an unweighted *k*-nearest neighbour method with OVA and OVO classifiers (*k*-NN OVA and *k*-NN OVO) were utilized in Publication V. It used the unweighted HVDM (Equation 2) as a distance measure. The basic idea of the OVA classifiers are described in Chapter 3.2.4 Attribute Weighted *k*-Nearest Neighbour Method with OVA classifiers.

In the OVO binarization, one classifier is learnt for each pair of classes. The number of classifiers is the number of all pairs of the classes (equation given in Figure 9). The *k*-NN OVO searches for the *k* most similar cases of a new case from each pair-wise classifier separately. Each classifier suggests (gives a vote) a class for the new case. The final class of the new case is chosen by the max-wins rule: A class gaining majority of the votes is set for the new case. If there are several class suggestions with the same number of votes (a tie situation), the class of the new case is solved by searching for the nearest case from the classifiers belonging to the tied classes and giving the class of the nearest case (the class with minimum distance) to the new case.

The classification runs were made with the *k*-NN implemented with Java (the *k*-NN included in the ONE). The classification runs were driven in Windows command prompt with batch files made by the author.

3.2.7 Support Vector Machine

The binary classifiers OVA and OVO (Figure 9) were applied also with support vector machines (SVM, also called maximum margin classifier) [Cortes and Vapnik, 1995] in Publication V. The SVM is originally developed to separate only two classes from each other, but, it has been extended for the multi-class classification tasks by constructing and combining multiple binary classifiers [Hsu, Lin, 2002; Joutsijoki, 2012]. Thus, it is necessary to use the OVA or the OVO classifiers with the SVM in order to utilize it in the multi-class classification. With the SVM OVA, a winner-takes-all rule was applied if there was only one classifier suggesting a class. If all OVA classifiers suggested the other class, then the basic 1-NN was applied to get the class for the new case. In the SVM OVO, each classifier suggested a class for the new

case. The final class was chosen by the max-wins rule: A class gaining majority of the votes was set for the new case. If there occurred a tie situation with the SVM OVO, the class was solved also with the 1-NN.

The SVM is a kernel-based classification method that generates a hyperplane (a linear decision function) to divide an input space so that the distance (the margin) between the separating hyperplane and the closest members of both classes is maximized [Cortes and Vapnik, 1995] (Figure 10). The closest members are called support vectors.

The binary SVM implementation of Bioinformatics Toolbox of the Matlab with the Least-Square method was used as a basis for the multi-class extensions. The SVM classification runs were made with linear, polynomial (d=2,3,4,5), Multilayer Perceptron (MLP) (scale κ in [0.2,10]; bias δ in [-10,-0.2]) and Gaussian Radial Basis Function (RBF) (scaling factor σ in [0.2,10] kernels with box constraints [0.2,10] (κ, δ and σ with intervals 0.2) with the Matlab by Henry Joutsijoki. The results of the best kernel functions (linear and RBF) were selected into comparison with the *k*-NN in Publication V.

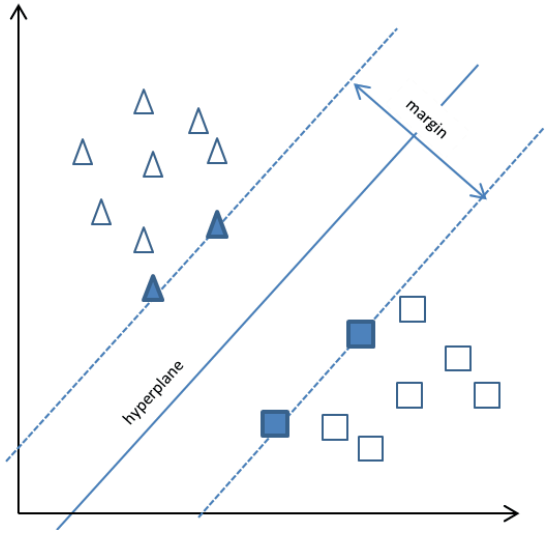


Figure 10. Support vector machines generate a hyperplane that separates two classes with the maximum margin

3.3 Knowledge Discovery Methods

3.3.1 Fitness Value Formation Method

In the fitness value formation for the attribute values, a machine learning method that learns the fitness values from the domain data was used [I; II]. The fitness value calculation method is based on the frequency distribution of the attribute values presented originally in [Viikki and Juhola, 2001]. It was shown to be useful in knowledge base refining. The fitness values are calculated separately for each attribute in different patterns from training cases belonging to the pattern. The fitness values can be defined as values that show how often values of an attribute occur in a certain pattern. The most frequently occurring attribute value fits the best for the pattern and, thus, for this attribute value is given the fitness value 100. If attribute value does not exist in the frequency distribution, its fitness value is set to 0. For the other attribute values, the fitness values are calculated by relating their frequencies to the frequency of the most frequently occurring value:

$$fv(c,a,v) = \frac{fr(c,a,v)}{fr(c,a,f)}100, \tag{7}$$

where *fv(c,a,v)* is the fitness value of the value *v* of the attribute *a* in the class *c*,
fr(c,a,v) is the frequency of the value *v* of the attribute *a* in the class *c* and
fr(c,a,f) is the highest frequency given by *f* in the distribution of the attribute *a* in the class *c*.

The machine learning method is presented in more detail and an example for forming the fitness values are given in Publication II.

The fitness values were calculated with the fitness value method in Publications I, II, III and IV and the effect of fitness values for the classification was examined in Publications I and II. The fitness value formation method was implemented with the Matlab by Kati Iltanen.

3.3.2 Scatter Method

The first knowledge discovery method applied in the attribute weighting in Publication III was the Scatter method [Siermala *et al.*, 2007; Juhola and Siermala, 2012]. The Scatter method can be utilized, for example, to evaluate the importance and separation power of attributes and to map the overlap of the classes in the

attribute space. It evaluates a measure for attribute separability and it has been applied in attribute subset selection.

The Scatter method is based on traversing through a data set by seeking the nearest unvisited case one at a time and concurrently counting the class changes between cases [Juhola and Siermala, 2012]. The nearest case is searched for with the Euclidean distance measure. If there are several cases with exactly the same distance, the nearest case is selected randomly from these nearest unvisited cases. A scatter value is computed based on the current number and theoretical maximum number of class changes during the data set traversing. The scatter value expresses the attributes' power to separate classes in the data set, the overlap of the classes within the attribute values: The closer the scatter value is to zero, the better the attribute differentiates the classes. The Scatter algorithm is described in more detail in [Juhola and Siermala, 2012; III].

The Scatter algorithm does not have any prerequisites for the class distributions. However, it needs complete input data to work properly and, therefore, the otoneurological data was necessary to be imputed before the scatter value calculation. In addition, the attribute values needed to be normalized into the same scale [0,1]. In Publication III, the scatter value was calculated for each attribute within each 'class versus all the other classes' situations in order to study each attribute's power to separate a class from the other classes. In addition, the scatter values were necessary to calculate from the whole training data set having the original classes in order to use the weights with the weighted 1-NN when all the *wk*-NN OVA classifiers voted a case to be a non-member.

In order to apply the scatter value as an attribute weight, it was necessary to take the inverse of it: The attribute weight value describes the opposite of the scatter value - the greater the attribute weight value is, the more important the attribute is.

The Scatter based weights were computed during research of Publication III and applied also in Publication IV. The scatter values were computed by the Markku Siermala and also with the Scatter method program[1] and the inversed scatter values were calculated with the Matlab by the author.

[1] You can find the Scatter method software by Markku Siermala and instructions how to use it from page http://www.uta.fi/sis/cis/research_groups/darg/publications.html. In the page, search for the year 2012 and find the second article (M. Juhola and M. Siermala: A scatter method for data and variable importance evaluation. *Integrated Computer-aided Engineering* 19, 137-149, 2012).

3.3.3 Instance-Based Learning Algorithms IB4 and IB1w

The second knowledge discovery method applied in the attribute weighting in Publication III was a weight calculation method from an incremental instance-based learning algorithm IB4 [Aha, 1992]. The IB4 is said to handle skewed class distributions and to tolerate irrelevant attributes by learning attribute relevancies (weights for attributes) independently for each class [Aha, 1992]. The attribute weights reflect the relative relevancies of the attributes in the class. In the IB4 method, each class is described with a separate class description and a set of attribute weights. A classification record (number of correct and incorrect classification predictions in the past) is maintained for each saved case in the class description. These classification records are used to detect noisy cases, statistically non-acceptable cases that are discarded from the class description. Statistically acceptable cases are used in the subsequent classification tasks. The weights are adjusted after the classification of each training case with a simple performance feedback algorithm that alters the class-wise weights based on the classification results: The weights of the attributes are increased when they correctly predict the classification and are otherwise decreased [Aha, 1992]. The feedback algorithm is presented in detail in [Aha, 1992; III].

The effect of irrelevant attributes on the classification is decreased by using class-wise attribute weights in a similarity function. Similarity is calculated with a negative attribute-weighted Euclidean distance measure. The attribute values are normalized to the range [0,1] in order to have the same (maximal) effect with each attribute. The IB4 can handle missing attribute values but, due to the Scatter method, the data was necessary to impute and to keep them comparable, the imputed data was utilized also with the IB4 weight calculation method and its variant IB1w. The IB1w is based on an IB1 method [Aha *et al.*, 1991] that combines the attribute weight algorithm from the IB4. The IB1 method usually handles all the classes at the same time with one classifier so, it was necessary to expand the IB1 to handle multiple disease classes. The weight calculation method from the IB4 was altered to the IB1w method by leaving out the case discarding, in other words, the IB1w keeps all cases in the class description during the process.

The attribute weights with the IB4 and the IB1w methods were computed in the research of Publication III and applied also in Publication IV. The IB4 and the IB1w were implemented with the Matlab by the author.

3.3.4 Genetic Algorithm

In the research of Kelly and Davis [Kelly and Davis, 1991], the results of the *k*-NN could be improved with the help of an adaptive weight calculation method. Thus, the attribute weighting was continued with evolutionary algorithms [Michalewicz, 1992] in order to study if the performance of the classification methods could be improved with an evolutionary algorithm adjusting the attribute weights. Genetic algorithms (GA) [Goldberg, 1989; Mitchell, 1996] and other evolution algorithms have been utilized in many different optimization and simulation tasks successfully because of their powerful search and optimization capabilities. The search method of the GA is a combination of directed and stochastic search and the search can be done multidirectionally because the GA maintain a population of potential solutions from the search space [Michalewicz, 1992].

In the beginning of the GA, a population of individuals is formed either randomly or with information about the domain. In each generation, the individuals are evaluated with an objective evaluation function that calculates each individual a fitness rate. Some individuals are randomly selected to reproduction where they either are crossovered or mutated. In the crossover, the information of two individuals is swapped in their corresponding elements. Mutations alter one or more elements of the individual arbitrarily. A selection method is utilized to find the fittest individuals for a new population. Elitism can be used in order to preserve the high-performance individual unchanged in the population [De Jong, 1975]. The GA ends after a fixed number of generations or if no further improvement is observed.

In Publication IV, the machine learnt weight sets of the Scatter, the IB4 and the IB1w methods together with three weight sets defined by the domain experts were utilized as a basis in the GA. From each of these weight sets were formed two different modifications by mutating the weight set with 50% probability and, in addition, three totally random individuals were generated in order to have a starting population consisting of 21 individuals. The weight sets were normalized before the GA runs. Each individual in the population contained real-valued weight sets for 94 attributes in seven disease classes. The GA runs were done separately with the ONE, the *wk*-NN OVA and the *cwk*-NN methods applied in population evaluation. With the ONE and the *cwk*-NN, the total classification accuracy was used as a fitness rate to an individual whereas, with the *wk*-NN OVA, the TPR of 7-NN (except with the disease class BRV that used 3-NN) was used as a fitness rate. The *wk*-NN OVA handled only one disease class (and its weight set) at a time whereas the ONE and the *cwk*-NN classified all seven disease classes at the same time.

A roulette-wheel selection was used in parent selection. The crossover was done in 80% probability and the crossover points were selected randomly and independently for each gene. Mutation was done in 1.0% probability for the gene: A random value was selected from the range [0, 1]. Elitism was applied in order to keep the best individual unchanged within the population during the evaluation. Otherwise, a survivor selection was utilized to maintain the population in 21 individuals in each generation: Individuals were ordered by their fitness rate and the individuals with the lowest fitness rates were discarded from the population. The GA ended after 20 generations or if the best fitness rate stayed the same during 10 successive generations. The GA with the ONE as the evalution method was tested also with 100 generations.

The more detailed description of the utilized GA and its pseudocode are presented in Publication IV. The GA and its evaluation methods were implemented with the Matlab by the author.

3.4 Evaluation

3.4.1 10-Fold Cross-Validation

The method of 10-fold cross-validation (CV) [Mitchell, 1997; Witten *et al.*, 2011] was used within the studies using the HUCH data with the machine learning methods to estimate the predictive classification performance of the methods. In the 10-fold cross-validation, each HUCH data subset was once utilized as the testing set while the other nine subsets formed the training set. The HUCH data set was randomly divided into 10 subsets of approximately equal size. The HUCH data division was made in a stratified manner to ensure that the class distribution of each subset resembled the skewed class distribution of the entire data set and that each subset contained data from each collection time. Thus, a stratified 10-fold cross-validation was applied. Therefore, the number of cases in different cross-validation subsets varied from 99 to 107 cases instead of 103 cases. The data subset divisions in the 10-fold cross-validation runs were little different within the studies, thus, having different training and testing sets used. In Publications III, IV and V, the data subset divisions were the same, thus, having comparable results to each other.

In Publication I, the 10-fold CV was repeated only once with the ONE using machine learnt fitness values due the implemention of the ONE of the time (manual

classification runs in the command prompt). The same 10-fold CV division was used with the fitness value formation method than with the classification method. In Publication II, the 10-fold CV was repeated three times with the ONE using machine learnt knowledge bases (KB2–KB5) and with the *k*-nearest neighbour method and the Naïve Bayes classifier. In Publication III, the 10-fold CV was repeated 10 times with the ONE and the attribute weighted *k*-nearest neighbour with OVA classifier (*wk*-NN OVA). Besides the classification methods, the cross-validation was taken into account also with the attribute weighting methods that repeated weighting runs ten times. In Publication IV, the GA runs using the nearest pattern method of the ONE and the attribute weighted *k*-nearest neighbour method using neighbour's class-based attribute weighting (*cwk*-NN) repeated the 10-fold CV 10 times. Instead, the GA runs using the *wk*-NN OVA as an evaluation method repeated the 10-fold CV only five times due to its huge computation time: The evaluation of one training data set (21 individuals) within one GA run using the *wk*-NN OVA in evaluation lasted from 5 h to 21 h depending on the used computer. In Publication V, the 10-fold CV was also repeated 10 times with the basic *k*-NN and the support vector machines and their variants using the OVA and the OVO classifiers

In the GA, the 10-fold cross-validation was applied a little bit different manner than with the other studies of the dissertation. During the GA runs, one cross-validation subset was left aside to be used in testing the individual having the highest fitness rate after the GA and nine cross-validation subsets were used during the GA runs. In the GA, six cross-validation subsets were used in training and three cross-validation subsets were used in testing the evaluation method. Thus, within the evaluation methods of the GA 60%–30% data division was used. The fitness values for the ONE were calculated from this 60% training data (six cross-validation subsets).

Because the TAUH data set contained only cases from four disease classes and its class distribution was badly skewed, we decided not to use the 10-fold CV with it. Instead, in Publication II, we used the TAUH data set as a testing set while using the whole HUCH data set as a training set. With the classification runs using only the machine learnt knowledge, the TAUH data set with 253 cases from four disease classes was used and with the classification runs using also the expert's knowledge, the TAUH data set with 228 cases from three disease classes was used in testing.

3.4.2 Evaluation Measures

The performance of the classification methods were evaluated mainly with two measurements: with a class-wise true positive rate (TPR) and a total classification accuracy (ACC). The TPR is calculated as a percentage of correctly inferred cases in the class:

$$TPR_c = 100\frac{t_{pos_c}}{n_{cases_c}}\%, \quad (8)$$

where t_{pos_c} is the number of correctly classified cases in the class c and
n_{cases_c} is the number of all cases in the class c.

The ACC describes the overall success rate of the classification method within the data set, it gives the percentage of all correctly classified cases in the data set:

$$ACC = 100\frac{t_{pos}}{n_{cases}}\%, \quad (9)$$

where t_{pos} is the total number of cases correctly classified in all classes and
n_{cases} is the total number of cases used in the classification.

The TPRs and ACCs presented in the result tables are given as mean values of the 10-fold cross-validations, except with the runs with the pure experts' knowledge.

Error bars (with 99% confidence intervals) were used in Publications II and III to show if there were significant differences with the results of used classification methods. Error bars were used to show the differences with the total classification accuracies and the median true positive rates of the three CV-runs between the first diagnosis suggestion of the ONE with different knowledge base combinations and between the classification methods (the ONE, the *k*-NN and the NB) using the augmented HUCH and TAUH data in Publication II. In Publication III, the error bars from the mean total classification accuracies, mean median TPRs and mean Cohen's kappa values from the 10 times repeated 10-fold cross-validation runs were shown with different 'machine learning - weight set' combinations.

In order to take into account also the effect of chance in the classification, a Cohen's kappa (K) [Cohen, 1960; Ben-David, 2007] (also called Kappa statistics [Witten *et al.*, 2011]) was utilized in the evaluation of the results in Publication III. The Cohen's kappa measures the agreement between predicted and observed classifications in the data set [Witten *et al.*, 2011], in this case, the pair-wise agreement between the classification method and human expert classification.

$$K = \frac{P_o - P_c}{1 - P_c}, \tag{10}$$

where P_o is the total agreement probability (*i.e.* the accuracy) and
P_c describes the 'agreement' probability that can be attributed to chance alone (kappa chance value) [Ben-David, 2007].

Kappa can get values from range [-1, 1], where -1 means total disagreement (worse than random performance), 0 means a random or majority-based classification and 1 means perfect agreement. When the kappa value is higher than 0.81, the pair is considered to have almost perfect agreement [Landis and Koch, 1977]. In Publication III, the Cohen's kappa was calculated separately for each 'classification method - weight set' combination to estimate the degree of agreement between their classification results and the actual class labels. It was also applied to evaluate the pair-wise agreement between the compared combinations. In addition to the kappa value, a probability of predicting the correct class due to chance (*Pc*) was presented in results of Publication III.

4 RESULTS

4.1 Publication I: Refinement of the Decision Support System

Publication I described the state of the otoneurological decision support system ONE after the upgrade and refinement process. A transformation of the ONE from C++ program running under MS-DOS to Java program was started in the beginning of the 2000s by Tapani [Tapani, 2008]. During the transformation, the program was modernized in many ways, for example, the graphical user interface was further developed to be more user-friendly (the navigation tree was added in the user interface) and another inference method (the *k*-NN) was added into the system. After the programming environment change, the answer database of the ONE was transferred from the Paradox database first into a basic text file and later by the author into a MySQL database. Before taking the MySQL database in use, it was checked that the paper data collection questionnaire and the query base of the ONE corresponded to each other (they utilized the same attributes). A few differences were noticed: Some questions were asked in the questionnaire but not in the ONE and vice versa. In addition, the otoneurological questionnaire was updated: A few totally new questions were added into the questionnaire and changes were made to answer alternatives of categorical questions to have the same amount of answer alternatives. Domain experts decided that a smaller amount of answer alternatives was adequate in describing the diseases. For the most of the categorical attributes this meant the decrease of answer alternatives. Decrease was done with the help of the domain experts by combining the answer alternatives together. Due to found differences and changes made to the questionnaire, the otoneurological paper questionnaire and the query base of the ONE were harmonized and it was necessary to go through the disease patterns in the knowledge base of the ONE with the otoneurological experts and update the weights and fitness values of the attribute values, if necessary. The updated questionnaire can be found from Appendix 1.

In addition to the manual update of the knowledge base with the experts, a machine learning method was utilized to refine the fitness values in the knowledge base. This fitness value formation method was based on the frequency distribution of the attributes [Viikki and Juhola, 2001]. The experimental results had shown its

usefulness in the knowledge base refining. The machine learning method computed the class-wise fitness values for the values of the attributes from the HUCH data. The fitness values were computed also for the binary type attributes (no/yes) whereas the experts' knowledge bases handled the binary attributes either existing (yes) or non-existing (no) (*i.e.* if the patient had answered "yes" to the current question, the weight of the attribute was added into the score calculation of the ONE and otherwise left unnoticed). Because some of the symptoms were not characteristic for the disease, there were attributes with negative weight values in the experts' knowledge base, thus, meaning that the attribute did not fit for the disease. Due to the fitness value addition into the binary attributes, it was possible to discard the negative attribute weight values and, instead, describe with the fitness values of the attribute values what was characteristic or uncharacteristic for the disease. For example, hearing loss is not usually related to vestibular neuritis, but, nevertheless, almost 20% of vestibular neuritis cases in this data set have hearing loss.

During the upgrade process, data transfer methods were developed in order to ease the data transfer into the decision support system and to ensure the quality of data. Previously, all answers were input manually into the system by clicking the given answer in the equivalent question in the system. Manual input was very time consuming, laborious and vulnerable for errors. Therefore, the paper questionnaire was altered to a scannable version. The harmonized questionnaire was changed from a basic Word document to a questionnaire in Snap Survey Software (Snap). The answers from the questionnaire were possible to read into the Snap with a text scanner. Naturally, someone still needed to check that the answers were read correctly into the system and, also, to input all free text fields by hand. Another solution for the data transfer was a web questionnaire that a patient could fill in beforehand [Mäkiranta, 2005]. With the web questionnaire, the patient information could have been in use for the ONE in real time because it used the same database as the ONE and, thus, the answers would have been in the correct form. Unfortunately, it was not yet then possible to take the web questionnaire in use due to the strict data security policies of the Finnish laws.

The classification runs were made with 951 cases from seven disease classes of the HUCH data with the ONE using the original knowledge base defined by the experts, the refined knowledge base defined by the experts, the knowledge base containing the machine learnt fitness values and the weights set to one and the knowledge base containing the machine learnt fitness values and the weights set by the experts. It was noticed after the results were published that the knowledge bases of the original and the updated experts' knowledge contained all 15 disease patterns

whereas the other knowledge bases contained patterns for nine disease classes. This made the classification a bit harder for the original and the refined expert knowledge bases.

The fitness value formation from the data for the values of the attributes seemed to work because the highest total classification accuracy (67.4%) when looking at the first diagnosis suggestion of the ONE was yielded with the knowledge base having the machine learnt fitness values with all weight values set to one. This knowledge base classified cases 27.0% better than the updated experts' knowledge base and 7.1% better than the knowledge base combining the experts' weights and the machine learnt fitness values. However, the combination of the machine learnt fitness values and the experts' weight values classified the cases best (91.8%) when looking the first, second and third diagnosis suggestions of the ONE. Its total classification accuracy was 9.0% better than the updated experts' knowledge base and 5.8% better than the machine learnt fitness values with the weights set to one.

The results of Publication I showed the benefits of the knowledge base refinement. The machine learnt fitness values improved the total classification accuracy compared to the pure domain experts' knowledge. However, there still seemed to be difficulties in disease separation. For example, BPV and BRV were recognized better with the refined pure experts' knowledge than with the knowledge base combining the machine learnt fitness values and the experts' weights (TPR of BPV was 56.7.% and BRV 70.0% with the refined expert knowledge and 50.3% and 50.0% with the knowledge base combining the experts' and ML knowledge). Also ANE cases was recognized worse with the knowledge combining the experts' attribute weights and the machine learnt fitness values: Only 16.8% of the cases were found as the first diagnosis suggestions and 71.8% when looking the three first diagnosis suggestions. ANE cases were found the best with the knowledge base containing the machine learnt fitness values and all weights set to one: 66.4% was found as the first and 87.0% with the three first suggestions.

4.2 Publication II: Machine Learning Method for the Fitness Value Formation

In Publication II, the machine learning method for the fitness value formation was presented and examined in more detail because the preliminary results of the method were promising in Publication I. First, the fitness value calculation method was utilized with the HUCH data with seven disease classes (951 cases). Five different

knowledge base combinations were tested: the refined pure experts' knowledge (KB1), the machine learnt fitness values with the weights set to one (KB2), the machine learnt fitness values with the experts' weights (KB3), the machine learnt fitness values with the experts' weights while the weights of ANE were set to one (KB4) and the machine learnt fitness values with the experts' weights while the weights of ANE and MEN were set to one (KB5). Second, the fitness values were calculated from the HUCH data set for three disease classes (BPV, MEN, VNE) and tested with the TAUH data set (228 cases). Third, the results of the nearest pattern method of the ONE with the knowledge base containing the machine learnt fitness values and the weights set to one were compared with the results of the *k*-nearest neighbour method and the Naïve Bayes classifier. The HUCH data with nine disease classes (1030 cases) was used. The methods were tested also with the TAUH data with four disease classes (253 cases) while the HUCH data was used as the training data. Due to the Naïve Bayes classifier, continuous attributes were discretized in all data sets used Publication II.

The results of the HUCH data with seven disease classes showed that the machine learnt fitness values for the attribute values improved the classification of the nearest pattern method of the ONE. The knowledge base containing the machine learnt fitness values with the weights set by the experts' (KB3) improved the total classification accuracy 14.9% compared with the pure experts' knowledge (KB1) when looking at the first diagnosis suggestion of the ONE. The total accuracy of the KB3 was 57.3%. However, the results showed that combining the fitness values and the attribute weights was difficult. When using the machine learnt fitness values and the experts' weights, the true positive rates (TPR) of disease classes decreased about for the half of the disease classes. Especially, acoustic neurinoma cases were recognized poorly (15.5%) with the machine learnt fitness values and the experts' weights (KB3) as the first diagnosis suggestion. Instead, cases of Menière's disease and vestibular neuritis were recognized better with the KB3 than with the pure experts' knowledge (KB1). When using the machine learnt fitness values for the attribute values with the weights set to one (KB2), the total classification accuracy improved almost 20% compared with the pure experts' knowledge (KB1) when looking at the first diagnosis suggestion of the ONE. The total classification accuracy of the KB2 was 62.1%. The total classification accuracies of the KB2 and the KB3 were closer to each other: The total accuracy of the KB2 was 4.8% better than with the KB3 as the first diagnosis suggestion. Acoustic neurinoma, Menière's disease and traumatic vertigo cases were recognized better with the weights set to one (KB2) than the weights set by the experts (KB3). Therefore, two variations of the

knowledge base KB3 were made. First, the attribute weights of acoustic neurinoma were set to one whereas the other diseases used the attribute weights defined by the experts' (KB4). Second, the attribute weights of Menière's disease were set to one besides the acoustic neurinoma (KB5). Interestingly, acoustic neurinoma cases were recognized the best (69.7%) when it was the only disease class with the weights set to one. When also the attribute weights of Menière's disease were one, the recognition of acoustic neurinoma decreased to 62.1% However, the cases of Menière's disease were recognized the best (91.5%) then. Also, the highest total classification accuracy (67.7%) was achieved when the weights of acoustic neurinoma and Menière's disease were set to one (KB5). However, when comparing the error bars with 99% confidence intervals for the total classification accuracies of each cross-validation within different knowledge bases, it could be seen that the accuracies did not differ significantly. Only the accuracy of the KB5 was significantly higher than with the KB3 in the cross-validation runs 1 and 2. Similarly, the error bars of the median TPRs did not reveal any significant differences.

The TAUH data with three disease classes was tested with the fitness values formed from the HUCH data. When looking at the first diagnosis suggestion, it seemed that the machine learnt fitness values improved the classification. With the experts' knowledge base, 23.2% of the cases were recognized whereas the other knowledge bases recognized from 42.5% to 53.5% of cases. However, when looking at the first, second and third diagnosis suggestions, the experts' knowledge base (KB1) yielded better results (total accuracy 71.9%) than the machine learnt fitness values with the weights set to one (KB2; total accuracy 59.2%). Still, better total classification accuracies (from 79.8% with the KB5 to 86.0% with the KB3) were achieved when the experts' weights were combined with the machine learnt fitness values. The results revealed that the knowledge transfer is not an easy task: The knowledge discovery from the data collected elsewhere does not necessarily solve the problem. Medical diagnosis is subjective and it can differ significantly depending on the physician doing it and even with the same person at different times [Kononenko *et al.*, 1998]. In additon, the HUCH and the TAUH data were collected with slightly different questionnaires at different institutes in Finland at different times. Unfortunately, the TAUH data set was so small and imbalanced that these results could be regarded only preliminary ones.

The classification results of the ONE, the *k*-NN and the NB were compared with each other with the HUCH data using nine disease classes and the TAUH data using four disease classes. The ONE had the lowest total classification accuracy with the HUCH data (60.2%) but it recognized three disease classes (SUD, VES, CL) better

than the other methods. The 5-nearest neighbour method (5-NN) had the highest total classification accuracy with both data sets (75.0% in the HUCH and 51.0% in the TAUH) and it had the highest true positive rates also with three disease classes (BPV, MEN, VNE) in the HUCH data. But, when looking at the error bars with the 99% confidence intervals for the total classification accuracies of each cross-validation made with the HUCH data, it could be seen that the 5-NN did not differ significantly from the 1-nearest neighbour method (1-NN) or the Naïve Bayes. On the contrary, the total accuracy of the ONE with the HUCH data was significantly lower than with the other methods. However, there were no significant differences in the median true positive rates of the methods based on the error bars. With the TAUH data, each classification method had the highest TPR with one of the four disease classes. The total classification accuracies with the TAUH data varied from 38.3% (1-NN) to 51.0% (5-NN), thus, expressing the difficulty of the domain.

In Publication II, a need for aid in the attribute weighting was shown. With the machine learnt fitness values for the values of the attributes it was possible to improve classification results but the fitness values alone were not adequate enough in separating classes from each other. Especially, BPV and BRV cases were confused with other diseases without weight values (equal weighting). The attribute weights set by the experts were preliminary tested with the machine learnt fitness values. However, this combination did not work well with all disease classes. The classification accuracy within the first diagnosis suggestion of the ONE was 4.8% better with the equal weighting than with the experts' weights but, when looking at the results of the two and three first diagnosis suggestions of the ONE, the experts' weights improved the classification accuracy compared with the equal weighting (6.3% and 6.8%, respectively). Interestingly, the best total classification accuracy was achieved with the knowledge base where the weights of the attributes in the classes acoustic neurinoma and Menière's disease were set to one and with the other classes were used the weights set by the experts (KB5). Its total classification accuracy was 67.7% when looking at the first diagnosis suggestions and 92.8% when looking at the three first diagnosis suggestions. Methods to find proper weights for the attributes to combine with the machine learnt fitness values for the attribute values are needed.

4.3 Publication III: Attribute Weighting with the Scatter and Instance-Based Learning Methods

Publication III concentrated on the class-wise attribute weight calculation. The class-wise attribute weighting was needed due to the inference engine of the ONE. The disease patterns are described separately in the knowledge base of the ONE and, thus, each disease class needs its own attribute weights. Therefore, it was necessary to use weighting methods that could learn weights for the attributes separately for each class and could express the relevance of a single attribute. The methods fulfilling these requirements were the Scatter method for the attribute importance evaluation [Juhola and Siermala, 2012; Siermala *et al.*, 2007] and the weight calculation method of the incremental instance-based learning algorithm IB4 [Aha, 1992]. The performance of the machine learnt attribute weights were compared with the performance of the weights set by the experts and the weights set to one (equal weighting). The attribute weighting was tested with two different classification methods, with the nearest pattern method of the ONE and with the attribute weighted *k*-nearest neighbour method using One-vs-All (OVA) classifiers (*wk*-NN OVA). The OVA classifiers were used with the weighted *k*-NN in order to keep the classification methods comparable to each other (*i.e.* the methods used the same attribute weights). The classification runs were made with the HUCH data using seven and nine disease classes. Seven disease classes were used within the runs where the results were compared with the experts' weights and nine disease classes within the runs using the machine learnt weights.

The highest classification accuracy (79.7%) with seven disease classes was yielded with the weighted 5-NN OVA using the Scatter based weights (wscat 5-NN OVA). It had also the highest Cohen's kappa value (0.73) and median of TPR (75.2%). The best total classification accuracy with the ONE (74.6%) was achieved also with the Scatter based weights (ONE wscat). Instead, with the 1-NN OVA, the highest total accuracy (74.7%) was achieved with the weights set by the experts (we 1-NN OVA). The total classification accuracy with the ONE using the pure experts' knowledge was 43.3% and with the machine learnt fitness values and the experts' weights 57.6%. The equal weighting worked well with the ONE because of the valid fitness values formed by the machine learning method in Publication II: The ONE with the weights set to one had the total classification accuracy 73.8%. However, with the *wk*-NN OVA, better results were achieved with the experts' weights than with the equal weighting: The 1-NN OVA classified cases 3.2% better with the experts' weights than with the equal weighting and the 5-NN OVA 2.7% better with the experts'

weights. With seven disease classes, the Scatter based weights improved the total classification accuracy compared with the weights set to one and the expert defined weights both with the ONE and the attribute weighted 5-NN OVA. The IB4 and IB1w weights worked better with the ONE than the experts' weights but the experts' weights worked better with the 1- and 5-NN OVA than the IB4 and IB1w weights.

Even though the ONE and the 5-NN OVA using the Scatter based weights and the 1-NN OVA using the weights set by the experts had the highest classification accuracies, they had the highest TPRs only with one or two disease classes when looking the TPRs method-wise: the ONE wscat with MEN (91.9%), the we 1-NN OVA with VNE (74.4%) and the wscat 5-NN OVA with SUD (84.3%) and TRA (86.6%). Otherwise, the weight sets and the methods achieving the highest TPRs varied; even the equal weighting yielded the highest TPRs with the *k*-NN OVA (95.4% of MEN cases were recognized with the 5-NN OVA using the equal weighting but then SUD cases were lost, TPR of SUD was then 29.4%). Especially, the TPR of SUD cases increased 54.9% (!) when using the Scatter based weights with the 5-NN OVA instead of the equal weighting, thus, SUD having TPR 84.3%. The attribute weighting affected also the results of the 1-NN OVA: 23.0% more of SUD cases and 17.5% more of TRA cases were found with the Scatter based attribute weights than with the equal weighting (TPR of SUD was 68.7% and TRA was 80.8%). With the ONE, the IB4 weights improved the recognition of traumatic vertigo cases 15.5% compared to the equal weighting, thus, having TPR of 94.5%.

The effect of the machine learnt fitness values on the classification of the ONE can be seen when comparing the results of the equal weighting (ONE w1) with the pure knowledge set by the experts (ONE experts). The TPRs of VNE, MEN and ANE cases increased notably (even 51.9%, 49.7% and 41.2%, respectively) when using the machine learnt fitness values. Also, TRA cases were recognized 11.9% better. However, the recognition of BRV and BPV cases decreased -28.5% and -11.2% with the equal weighting. When adding different weight sets with the machine learnt fitness values, the results stayed quite near the results of the equal weighting. Except, when the machine learnt fitness values were combined with the experts' weights, ANE cases were totally lost: Only 16.7% of ANE cases were classified correctly as the first diagnosis suggestion with the ONE we.

The best total classification accuracy (73.3%) with nine disease classes was achieved with the weighted 5-NN OVA using the Scatter based weights, like with seven disease classes. It had also the highest kappa value (0.66) but the highest median of TPR was achieved with the ONE using the IB4 weights (65.7%). However, the best total accuracy with the ONE (62.4%) was achieved with the

Scatter based weights. With all methods (the ONE, the 1- and the 5-NN OVA), the best total accuracy was achieved with the Scatter based weights. The equal weighting achieved better results than the IB4 and the IB1w weighting with the ONE and with the 5-NN OVA. With the 1-NN OVA, the IB4 and the IB1w weights had similar level total classification accuracies with the equal weighting.

With the nine disease classes, the highest median TPRs were achieved with the ONE using the Scatter based weights only for one disease class (VNE 64.9%). The Scatter based weights yielded best TPRs for four classes with the 1-NN OVA (TRA 73.8%, VNE 65.8%, SUD 61.5% and VES 36.9%) and for four classes with the 5-NN OVA (TRA 85.2%, SUD 81.5%, VNE 75.4% and BPV 65.1%). Interestingly, the IB4 weights worked well with the ONE: 95.6% of TRA cases (26.0% better than the ONE w1), 89.2% of CL cases (42.5% better than the ONE w1!), 76.2% of SUD cases (14.9% better than the ONE1 w1) and 47.5% of VES cases (7.3% better than the ONE w1) were classified correctly. Instead, it did not recognize well cases of MEN (-15.6% compared to the ONE w1) and BPV (-7.5% compared to the ONE w1) why its total classification accuracy was lower than with the equal weighting or the Scatter based weights. The ONE with the Scatter based weights improved also the recognition of SUD and TRA cases compared to the equal weighting (increase 7.6% and 6.7%, respectively). The 1-NN OVA with the IB1 weights had the highest median TPR 86.0% with MEN (7.6% better than with the equal weighting).

The addition of two difficult disease classes into the classification (VES and CL) decreased the true positive rates of other seven disease classes. The effect of addition can be noticed especially from the disease class BRV, whose recognition decreased dramatically after adding vestibulopatia and central lesion into the knowledge base of the ONE. With seven disease classes, different knowledge base combinations with the ONE found correctly 23.5–65.0% of BRV cases but with nine disease classes only 3.0–20.5% of the cases. However, there were only 20 cases in the class of BRV and, thus, wrong classification of one case had a big influence on the TPR. The TPR of BPV and MEN cases decreased especially with the ONE after adding two disease classes into classification: 47.8–65.9% of BPV cases and 42.0–91.9% of MEN cases were found with seven diseases classes whereas with nine disease classes the TPRs varied with BPV 25.1–32.6% and with MEN 65.7–81.3%. Small decrease in the TPRs after the two class addition can be seen also in the results of the 1- and 5-NN OVA.

The mean confusion matrices of the ONE, the 1-NN and the 5-NN OVA showed that with the ML methods using machine learnt knowledge all disease classes were mixed up with Menière's disease both with seven and nine disease classes,

especially sudden deafness (from 11.3% to 53.6%), central lesion (from 16.7% to 44.2%), vestibulopatia (from 21.8% to 34.5%) and acoustic neurinoma (from 19.7% to 34.0%). Instead, the ONE with the knowledge base purely defined by the experts misclassified cases from the seven classes mainly as BRV (47.1% of VNE cases, 30.9% of MEN cases and 20.8% of BPV cases) or BPV (30.0% of BRV cases, 24.8% of VNE cases and 24.7% of TRA cases). However, ANE cases were misclassified with the ONE using the experts' knowledge to SUD cases (48.1%) whereas other methods confused it to Menière's disease. BRV cases were mixed up with VES cases (from 28.0% to 62.0%), BPV cases (from 9.0% to 44.5%) and MEN cases (from 12.0% to 24.0%) with the machine learnt knowledge bases. With the nearest neighbour methods using nine disease classes, CL and VES were misclassified also to BPV cases (from 21.7% to 31.2% and from 21.5% to 28.5%, respectively) besides MEN. With all ML methods, CL cases were also mixed up with VES cases (from 19.2% to 28.3%) and some VES cases to CL cases (from 4.4% to 23.1%).

As Publication III showed, the attribute weighting is demanding. The extent of the effect the attribute weighting had on the classification depended on the used classification method and on the disease classes to be classified. The Scatter based weights improved the total classification accuracies and median true positive rates compared to the equal weighting with the nearest pattern method of the ONE and the *wk*-NN OVA both with seven and nine disease classes whereas the IB4 and IB1w weights had a slight decreasing effect on the total classification accuracies. With the ONE, the machine learnt attribute weights yielded better classification accuracies than the attribute weights defined by the experts whereas with the *wk*-NN OVA the weights defined by the experts worked well. To find the right combination of the attribute weights and the fitness values for the attributes is a difficult task.

4.4 Publication IV: Genetic Algorithm Based Attribute Weighting

In Publication IV, the genetic algorithm (GA) [Mitchell, 1996] was utilized in the evolution of the attribute weight values. The machine learnt weight sets gained as results with the Scatter, the IB4 and the IB1w methods in Publication III in addition to the weight sets defined by the experts, modifications of these weight sets and few random weight sets were utilized as a starting point in the GA runs, thus, forming a population containing at total 21 individuals. Each individual in the population contained a real-valued weight sets for 94 attributes in seven disease classes.

The evaluation of the individuals in the population was made with three different methods separately. The individuals were evaluated with the nearest pattern method of the ONE, with the *wk*-NN OVA and with the attribute weighted *k*-nearest neighbour method using neighbour's class-based attribute weighting (*cwk*-NN). The *cwk*-NN used class-dependent weights with one classifier. Thus, it was possible to use and modify all weight sets at the same time with the *cwk*-NN. The evaluation methods calculated for each individual a fitness rate: With the ONE and the *cwk*-NN the fitness rate was the individual's total classification accuracy and with the *wk*-NN OVA the true positive rate of the individual with 7-NN (except with the disease class BRV was used the TPR of 3-NN due to its small number of cases). During the GA runs, elitism [De Jong, 1975] was applied. The current fittest individual was kept unchanged in the population in order to avoid missing the high-performance individual during the evolution. The other individuals were exposed to the roulette-wheel selection (selection of individuals into a mating pool), crossover (offspring creation, 80.0% probability) and mutation (1.0% probability).

The GA runs were made with the HUCH data using seven disease classes. The highest total classification accuracy (79.1%) and median true positive rate (73.6%) were achieved with the GA using the weighted 5-NN OVA as population evaluation method. Also, the weighted 1-NN OVA worked quite well with the GA: Its total classification accuracy was 76.2% and median TPR 71.5%. The GA using the ONE as the evaluation method yielded a bit lower results: Its classification accuracy was 73.8% and median TPR 66.2%. However, the GA ONE worked better than the GA *cwk*-NN: It classified only 61.1% of the cases correctly with the cw1-NN and 67.4% with the cw5-NN. The GA weighted 5-NN had the highest TPRs with three disease classes (MEN 92.2%, VNE 79.2% and SUD 77.0%), the GA weighted 1-NN OVA with two disease classes (BPV 74,1% and ANE 71.5%) and the GA ONE with two disease classes (TRA 83.0% and BRV 31.5%). Sudden deafness cases were badly lost with the GA *cwk*-NN, the GA cw5-NN recognized only 23.2% of the cases.

The best improvement between the starting and the ending population within 20 generations lasting GA was achieved with the GA using the *cwk*-NN in the evaluation. The best total accuracy in the starting population was 63.6% and in the ending population 68.3% whereas the worst total accuracies were 27.9% and 56.2%, respectively. The worst mean accuracies increased also within the GA ONE (from 49.8% to 61.4%) and the GA *wk*-NN OVA (from 75.3% to 78.7%). Otherwise, the effect of the GA generated weights on the classification was quite small and in some cases even decreasing with these two methods. With the *cwk*-NN, the GA improved the recognition of especially two disease classes, acoustic neurinoma and traumatic

vertigo. In the beginning, less than 50 % of ANE cases and less than 60% of TRA cases were found, but, after the GA runs, almost 71% of ANE cases and 72% of TRA cases were recognized correctly. The best total classification accuracy in the starting population (79.6%) within the evaluation methods was achieved with the weighted 5-NN OVA and in the ending population (79.5%) with the weighted 3-NN OVA.

The GA runs ended if the evolution lasted 20 generations or the highest fitness rate did not change during 10 generations. Most of the GA runs using the *cwk*-NN (82.0%) as an evaluation method lasted 20 generations whereas most of the GA runs using the ONE (75.0%) or the weighted *k*-NN with the OVA classifiers (82.9%) as the evaluation method ended before the 20th generation due to having the fitness rate unchanged for 10 rounds. Interestingly, the GA runs with the ONE ended after 10 rounds in 48.0% and with the *wk*-NN OVA in 54.9% of the runs which meant that there were no changes in the highest fitness rate during the GA run. Probably the reason for this was the elitism. All cross-validation runs of the disease class traumatic vertigo and most of the CV runs of sudden deafness (96.0%) and benign recurrent vertigo (94.0%) ended after 10th round during the GA runs using the *wk*-NN OVA in the evaluation. This explains why there were no big changes in the mean best total classification accuracies between the starting and the ending population with the ONE or the *wk*-NN OVA utilized in the evaluation. As a matter of fact, the best mean accuracy of the population did not change even if the GA runs with the ONE were run 100 times. Within the 100 round GA, the run was ended if the fitness rate did not change during 50 generations. This happened during 12.0% of the runs. Otherwise, 39.0% of the GA runs ended before the 100th generation.

In Publication IV, the genetic algorithm was utilized in the evolution of attribute weighting. The starting point was the weights of the attributes set by the domain experts and computed with the machine learning methods from the domain data and their mutations. Majority of the weight sets in the starting population were based on the domain knowledge and data, and, thus, were already more or less optimized on the problem at hand. Therefore, all the GA generated result weight sets did not improve the classification results.

4.5 Publication V: Multi-Class Classification Task Redefinition into Multiple Binary Problems

The effect of splitting a multi-class classification task into multiple binary classification tasks was examined in Publication V. Two commonly used approaches for splitting a multi-class task into multiple binary tasks, One-vs-All other (OVA) [Rifkin and Klautau, 2004] and One-vs-One (OVO) [Fürnkranz, 2001] classifiers were used. In experimental research of Fürnkranz [Fürnkranz, 2001], the use of OVO classifiers yielded significant improvements in the predictive accuracy compared with the OVA classifiers. The OVA and the OVO classifiers were tested with the basic unweighted *k*-NN and support vector machines (SVM) [Cortes and Vapnik, 1995] with different kernels. The basic *k*-NN was used as a baseline because it does multi-class classification with one classifier. In addition, it had the highest total classification accuracy in Publication II.

The HUCH data with nine disease classes were used in the classification. Due to the calculation of the SVM, the imputed data set was used with all classification methods in Publication V. With nine deducible disease classes, there were 9 class-wise classifiers in use with the OVA classifiers and 36 pair-wise classifiers with the OVO classifiers (Figure 9). The results of the best kernels (linear and RBF) of the SVM were reported in the publication besides the 1-NN and 5-NN methods.

The results of Publication V supported Fürnkranz's [2001] observation: The 5-nearest neighbour method using the OVO classifiers (82.4%) and the SVM with linear kernel using the OVO classifiers (77.4%) yielded better classification accuracies compared with results of the OVA classifiers (5-NN OVA 78.8% and SVM linear OVA 76.8%). The 5-NN with the OVO classifiers yielded also better classification accuracy than the basic 5-NN with the multi-class classifier (79.8%). However, the OVA classifiers worked a little bit better (1.2%) than the OVO classifiers with the SVM using the RBF kernel.

The 5-NN with the OVO classifiers had the highest TPR with six disease classes. Especially, the sudden deafness (SUD) was recognized better with the multiple binary classifiers than with one multi-class classifier. The basic 5-NN classified 77.5% of SUD cases correctly whereas the 5-NN OVA classified 87.5% and the 5-NN OVO 94.3% of the cases correctly. Traumatic vertigo (TRA) cases were recognized best with the SVM linear using OVO classifiers (99.9%). The 5-NN OVO recognized 96.2% of TRA cases, which was better than with the basic 5-NN (89.6%). Instead, the OVA classifiers decreased the recognition of traumatic vertigo both with the 5-NN (to 77.7%) and the SVM linear (to 79.9%).

The mean percentage of the occurring tie situations within the OVO classifiers was notably smaller than within the OVA classifiers. For example, the 5-NN using the OVO classifiers had tie situations only with 2.0% of cases whereas the 5-NN with the OVA classifiers had tie situations with 16.2% of cases. All the tie situations with the 5-NN OVA occurred when the classifiers voted the case to be a non-member of the classifiers. Instead, the tie situations with the 5-NN OVO occurred mainly with two or three disease classes: The cases of benign positional vertigo, Menière's disease and vestibulopatia were difficult to distinguish from each other because of their similar kinds of symptoms.

In this study, the attribute weighting was left aside in order to see the effect of multi-class classification task redefinition into multiple binary classification tasks on the classification results. Publication V showed that the use of the OVO classifiers improved the classification accuracies both with the 5-nearest neighbour method and the support vector machines using linear kernel.

4.6 Results Comparison

Because the data set divisions and validation methods in Publications I and II differed from the methods utilized in Publications III, IV and V, it is not possible to compare their results directly to each other. The continuous attributes were discretized in [I] and [II]. Furthermore, the different number of disease classes was used within the classification runs of [IV] and [V] and, therefore, only the results of [III] and [IV] and the results of [III] and [V] can be compared to each other. Within the research of Publication III was repeated the classification runs with the nearest pattern method of the ONE using the refined pure experts' knowledge (ONE1 experts in the [III] corresponds to KB1 in [II]), the machine learnt fitness values with the weights set to one (ONE1 w1 in [III] corresponds to KB2 in [II]) and the machine learnt fitness values with the weights defined by the domain experts (ONE1 we in [III] corresponds to KB3 in [II]) and, thus, the effect of methods presented in [II] can be taken into comparison through results in [III].

The best classification results of seven disease classes within the classification methods from Publications III and IV in addition to the results of three equivalent classification runs of Publication II are shown in Table 2. The highest total classification accuracy (79.7%) and the highest median true positive rate (75.2%) were achieved with the 5-NN using the Scatter based attribute weights and the OVA

Table 2. The best classification results within classification methods in Publications III and IV. Seven disease classes utilized in the classification. Results of ONE1 experts, ONE1 w1 and ONE1 we equivalent to methods in Publication II are added for comparison.

		[III]:						[IV]:		
Disease	Cases	ONE1 experts	ONE1 w1	ONE1 we	ONE1 wscat	we 1-NN OVA	wscat 5-NN OVA	GA ONE1	GA cw5-NN	GA w5-NN OVA
ANE	131	24.4	65.6	16.7	62.3	67.6	63.1	63.5	70.5	70.4
BPV	173	65.9	54.7	47.8	55.6	69.3	70.9	55.4	56.1	73.6
MEN	350	42.0	91.7	75.8	91.9	87.2	93.7	90.8	81.5	92.2
SUD	47	68.1	62.6	85.5	71.9	45.3	84.3	66.2	23.2	77.0
TRA	73	67.1	79.0	40.1	83.2	74.8	86.6	83.0	71.9	63.6
VNE	157	15.9	67.8	66.1	67.8	74.4	75.2	68.0	63.8	79.2
BRV	20	65.0	36.5	23.5	43.0	19.0	18.5	31.5	12.0	14.0
Median of TPR		65.0	65.6	47.8	67.8	69.3	**75.2**	66.2	63.8	**73.6**
Total ACC	951	43.3	73.8	57.6	74.6	74.7	**79.7**	73.8	67.4	**79.1**

classifiers (wscat 5-NN OVA). The highest total classification accuracy with the ONE (74.6%) was achieved also with the Scatter based weights.

In the first phase of the study, it was concentrated on the knowledge base refinement and the fitness value formation. The effect of the fitness value formation for the attribute values can be seen from the results of the ONE with the refined pure experts's knowledge (ONE1 experts) and from the ONE using machine learnt fitness values and weights set to one (ONE1 w1). The classification accuracy increased notably (30.5%) when using the machine learnt fitness values in the knowledge base of the ONE. In addition, cases of VNE, MEN and ANE were recognized much better when using the machine learnt fitness values (TPR increased 51.9%, 49.7% and 41.2%, respectively). However, BRV and BPV were recognized better with the pure experts' knowledge (28.5% and 11.2% better, respectively). Notice that BRV was a small disease class (only 20 cases) and, thus, one case had quite big influence on the TPR. The knowledge discovery method formed the fitness values also for the binary type attributes, which made it possible to take into account confounding symptoms of the cases.

The combination of the machine learnt fitness values and the weights set by the domain experts (ONE1 we) improved the classification accuracy (14.3%) compared to the ONE with the pure experts' knowledge. However, compared to the ONE1 w1, the total classification accuracy of the ONE we was -16.2% lower. Only SUD cases were recognized better with the ONE1 we than with the ONE1 experts (increase 17.4.%) or the ONE w1 (increase 22.9%) but, instead, ANE, BRV and

TRA cases were lost with the ONE1 we: Only 16.7% of ANE, 23.5% of BPV and 40.1% of TRA cases were recognized.

In the second phase, the effect of the machine learnt attribute weights was examined. The Scatter based attribute weights yielded a bit higher total classification accuracies than the equal weighting with the ONE (0.8% higher) and with the 1- and 5-NN OVA (1.3% and 3.5% higher, respectively). The TPRs of SUD and TRA increased with all three methods when using the Scatter based weights instead of the equal weighting, especially with the 5-NN OVA (SUD increased 54.9% (!) and TRA 18.7%). Otherwise, the machine learnt attribute weights had a slight decreasing effect on the classification: With the equal weighting in the *wk*-NN OVA was achieved a little bit higher classification accuracies (the total accuracy was 71.5% with the w1 1-NN OVA and 76.2% with the w1 5-NN OVA) than with the IB4 and the IB1w weights [III]. Also, with the ONE the IB4 weights decreased the total accuracy -3.8% compared to the equal weighting whereas the IB1w weights had almost the same total accuracy (73.9%) as with the equal weighting. Interestingly, the ONE with the IB4 weights improved the TPR of TRA 15.5% compared to the equal weighting, thus, recognizing 94.5% of TRA cases correctly. Instead, with the IB1w weights, TRA cases were lost both with the 1- and 5-NN OVA compared to the equal weighting (-12.8% and -13.5%, respectively). With the 1-NN OVA, TPR of SUD decreased both with the IB4 (-17.6%) and the IB1w (-18.3%) weights.

With the weights defined by the domain experts a bit higher classification accuracies were achieved than with the equal weighting with the 1- and 5-NN OVA methods (3.2% and 2.7%, respectively) whereas with the ONE, the total classification accuracy decreased -16.2% with the experts' weights. The machine learnt weights improved the total classification accuracies of the ONE (the Scatter based weights 17.0%, the IB4 12.4% and the IB1w 16.3%) compared to the ONE1 we whereas with the 1-NN OVA the best result was achieved with weights set by the experts. Only the Scatter based weights improved the total accuracy of the 5-NN OVA compared to the experts' weights. Still, the ONE using the weights set by the experts found SUD cases better than the ONE using the machine learnt weights whereas with the 1- and 5-NN OVA the Scatter based weights improved the TPR of SUD (23.4% and 32.4%) compared to the experts' weights.

The evolutionary approach was applied in the attribute weighting [IV]. The weight sets generated with the ML methods in Publication III were used in the starting population of the genetic algorithm. The populations were evaluated within the GA with the ONE (GA ONE), with the *wk*-NN using the OVA classifiers (GA *wk*-NN OVA) and with the attribute weighted *k*-NN using the neighbour's class-

based attribute weighting (GA *cwk*-NN). The GA ONE1 had the same total classification accuracy than the ONE with the equal weighting and, thus, it was a little weaker than the ONE with the Scatter based weights. The total classification accuracy of the GA weighted 5-NN OVA did not improve either compared to the wscat 5-NN OVA, but the use of the GA generated weights improved a bit the recognition of ANE (7.3%) and VNE cases (4.0%) but, instead, it lost TRA cases (TPR decreased -23.0%). The GA 1-NN OVA improved the total classification accuracy a bit (1.5%) compared to the 1-NN OVA we and, especially, the TPR of SUD (21.9%). With the GA *cwk*-NN was possible to test the evolution of all weights at the same time, but this classification method did not achieve good results: The total classification accuracies of the GA cw1-NN was 61.1% and with the GA cw5-NN 67.4%. Thus, it had better classification result than the ONE with the pure experts' knowledge or the ONE with the machine learnt fitness values and weights set by the experts but lower than the ONE with the equal weighting or the Scatter based weights. Only the TPR of ANE increased (7.9%) with the GA cw1-NN compared to the ONE1 wscat, otherwise the TPRs decreased (from -5.6% to -41.3%.

The best classification results of nine disease classes within the classification methods from Publications III and V in addition to the result ONE w1 equivalent of Publication II are shown in Table 3. The highest classification accuracy (82.4%) and the highest median true positive rate (88.2%) were achieved with the unweighted 5-NN OVO. Overall, the total classification accuracies and median of TPRs achieved in [V] were higher than the ones in [III]. However, the imputed data was utilized within the classification runs in [V] whereas in [III], the imputed data were used only in the attribute weight calculation, not during the classification. Therefore, the classification results of [V] are not after all directly comparable to the results of [III] even though they have used the same cross-validation sets in data. Nevertheless, the cases of VES and CL were recognized notably better with the ONE (with all weight sets) using the data with missing values than with the 5-NN or the SVM with different classifiers using the imputed data. Especially, the ONE IB4 recognized CL cases: Even 89.2% of the cases were classified correctly. In addition, the ONE IB4 recognized better VES (47.5%) and well TRA cases (95.6%). BRV was difficult for all the methods.

The second phase analysis of the effect of machine learnt attribute weights on the total classifcation accuracies was possible to do also with the nine disease classes. The machine learnt attribute weights affected the total classification accuracies quite similarly with the nine disease classes as they affected seven disease classes.

Table 3. The best classification results within different methods in Publications III and V. Nine disease classes utilized in the classification. Result of ONE1 w1 equivalent to method in Publication II is added for comparison.

		[III]:				[V]:				
Disease	Cases	ONE1 w1	ONE1 wscat	wscat 1-NN OVA	wscat 5-NN OVA	5-NN	5-NN OVA	SVM RBF OVA	5-NN OVO	SVM RBF OVO
ANE	131	65.6	62.7	60.6	60.0	89.5	90.2	90.7	95.0	87.2
BPV	173	32.6	31.4	57.5	65.1	77.9	77.6	78.6	79.0	67.0
MEN	350	81.3	80.2	77.3	93.1	92.4	89.8	91.5	93.1	90.1
SUD	47	61.3	68.9	61.5	81.5	77.5	87.4	58.1	94.3	79.4
TRA	73	69.6	77.3	73.8	85.2	89.6	77.7	96.7	96.2	99.3
VNE	157	63.9	64.9	65.8	75.4	87.7	85.0	84.3	88.2	81.4
BRV	20	4.0	4.0	19.0	14.5	3.0	8.0	8.0	4.0	16.5
VES	55	40.2	41.3	36.9	26.7	9.6	15.8	13.5	14.0	22.8
CL	24	46.7	46.7	22.9	7.5	5.0	15.0	15.8	2.1	28.5
Median of TPR		61.3	62.7	60.6	**65.1**	77.9	77.7	78.6	**88.2**	79.4
Total ACC	1030	62.2	62.4	64.6	**73.3**	79.8	78.8	79.4	**82.4**	78.2

The Scatter based weights slightly improved the classification accuracies compared to the 1-NN OVA (1.7%) and 5-NN OVA (3.2%) with the equal weighting whereas the results with the ONE were almost the same (difference only 0.2%) [III]. The TPRs of SUD and TRA increased again with all three methods when using the Scatter based weights instead of the equal weighting, especially SUD (53.2%) with the 5-NN OVA and TRA (14.6%) with the 1-NN OVA.

Compared to the equal weighting, the IB4 and the IB1w weights decreased the total classification accuracy of the ONE (-3.1% and -0.3%, respectively) and the 5-NN OVA (-0.9% and -3.5%, respectively). They did not affect the total classification accuracy of the 1-NN OVA. However, with the ONE, the IB4 weights improved the TPR of CL 42.5%, TRA 26.0% and SUD 14.9% and decreased the TPR of MEN -15.6% and the IB1w weights improved the TPR of TRA 9.7%. Again, TPR of TRA decreased both with the 1- and 5-NN OVA using the IB1w weights (-11.8% and -17.9%, respectively) and TPR of SUD decreased with the 1-NN OVA using the IB4 (-16.8%) and the IB1w (-13.0%) weights compared to the equal weighting.

In the third phase was concentrated on the classification task redefinition. The multi-class classification task was separated into several binary classification tasks. When comparing the different variations of the unweighted 5-nearest neighbour method, the OVO classifiers yielded the highest total accuracy (82.4%) and the

median TPR (88.2%). Interestingly, the 5-NN OVA yielded a bit lower total classification accuracy (78.8%) than the basic 5-NN (79.8%). As can be seen from the Table 3, it depends on the used classification method, which classifier combination worked the best. With the unweighted 5-NN and the SVM using linear kernel, the OVO classifiers worked the best but with the SVM using RBF kernel the OVA classifiers separated the cases the best. However, the difference was quite moderate between the classifiers.

The total classification accuracies decreased with the ONE and the weighted *k*-NN with the OVA classifiers when using nine disease classes instead of seven classes in the classification (Tables 2 and 3). With the Scatter based weights, the decrease in the total accuracies was -12.2% with the ONE, -8.2% with the 1-NN OVA and -6.4% with the 5-NN OVA. Addition of two difficult disease classes with similar kind of symptoms decreased the classification of all disease classes. Especially, with the ONE, the TPRs decreased the most with the disease classes BRV (from -22.5% to -39.0%), BPV (from -22.1% to -25.2%) and MEN (from -10.4% to -15.7%). With the *wk*-NN OVA, the decrease was lower: The worst decreases in TPRs were with BPV (from -9.8% to -12.2% with the 1-NN OVA and from -5.8% to -8.5% with the 5-NN OVA) and with the 1-NN OVA in TPRs of BRV (from -6.0% to -7.0%), SUD (from -3.4% to -8.7%), TRA (from -3.1% to 7.0%) and VNE (from -4.1% to -6.1%).

5 DISCUSSION AND CONCLUSIONS

In the dissertation, the machine learning methods were applied to the knowledge discovery in the otoneurological domain in order to refine the knowledge of the decision support system and to improve the classification accuracy of the system in real world situations. The phases of the dissertation were divided into three parts: fitness value formation for the attribute values, attribute weighting and classification task redefinition (Figure 1). In the first phase, it was concentrated on the knowledge refinement of the ONE with the domain experts [I] and on knowledge discovery method forming the fitness values for the values of the attributes [II]. The knowledge base of the ONE needed refining due to the harmonization of the otoneurological paper data collection questionnaire and the decision support system and due to changes made to the questionnaire, for example, changes made to answer alternatives of categorical questions. In the beginning, the refinement of the knowledge base was made manually with the domain experts and later with the machine learning methods. In Publication I, the effects of the manual refinement of the knowledge base made with the domain experts were examined and the fitness value formation method was preliminary tested with the otoneurological data. The refinement process enhanced the decision support system and its classification results. In Publication II, the fitness value formation method was presented in detail and the classification performance of the nearest pattern method of the ONE using the machine learnt fitness values was compared to the performances of the *k*-nearest neighbour method and the Naïve Bayes method. The machine learnt fitness values improved the classification accuracy of the ONE, which inference mechanism was shown comparable to the *k*-nearest neighbour and Naïve Bayes methods.

The second phase concentrated on the attribute weighting. In Publication III, the attribute weights were calculated with three different machine learning methods: with the Scatter, the IB4 and the IB1w methods. These machine learnt attribute weights were tested with the nearest pattern method of the ONE and the attribute weighted *k*-NN using the OVA classifiers (*w*/*k*-NN OVA). Also, the weights defined by the domain experts and equal weighting (all weights set to one) were tested with the classification methods. The extent how much attribute weighting affected the classification results depended on the used classification method and the disease

classes to be classified. With the nearest pattern method of the ONE, the machine learnt attribute weights yielded better classification accuracies than the knowledge base using the attribute weights and fitness values set by the experts. The Scatter based weights improved the classification accuracy compared to the equal weighting both with the ONE and the *wk*-NN OVA when using seven or nine disease classes.

The result weight sets of the ML methods from Publication III and the weight sets defined by the domain experts in addition their mutations and three random weight sets were utilized as a starting point in the evolutionary approach on the attribute weighting in Publication IV. The populations (the weight sets) were evaluated within the GA with the nearest pattern method of the ONE, with the attribute weighted *k*-NN using the neighbour's class-based attribute weighting and with the *wk*-NN using the OVA classifiers. The genetic algorithm approach in the attribute weighting did not improve the classification results as hoped. The total classification accuracies with the weight sets generated by the GA were quite near the results of the ONE and the weighted *k*-NN OVA with the machine learnt weight sets in Publication III. Only with the weighted 1-NN OVA, the GA generated weights improved the total classification accuracy a bit. The best classification accuracies within the starting and the ending population were quite near each other with different GA evaluation methods, but, instead, with the weights generated by the GA the worst total accuracies in the ending population compared to the starting population were improved. With the otoneurological data, the GA approach in attribute weighting did not affect the classification results so much maybe due to used weight sets in the starting population. The weight sets utilized in the starting population were based on the domain knowledge (the weight sets defined by the domain experts) and data (the machine learnt weights), and, thus, they were already more or less optimized on the problem at hand. With a totally random weight sets in the starting population improvement might have been more obvious after the GA.

In the third phase, the effect of the classification task redefinition was tested by separating the multi-class classification task into multiple binary classification tasks [V]. The attribute weighting was left aside in order to see the effect of classification task redefinition on the classification results. During the research, the OVA and OVO classifiers were utilized with the unweighted *k*-NN and with the support vector machines using different kernels. The results showed that the use of the OVO classifiers improved the classification accuracies both with the 5-nearest neighbour method and the support vector machines using linear kernel. Thus, the OVO classifier approach is worth testing also with the attribute weighting and with other

machine learning methods. In the future, the OVA and the OVO classifiers should be experimented also with the nearest pattern method of the ONE.

The results of the whole study support the statement of Kentala [Kentala, 1996b] that different otoneurological diseases can be challenging to differentiate from each other due to similar kind and overlapping symptoms. The value distributions of the occurrence and duration of the vertigo attacks and type of hearing loss (Figure 2, 3 and 7) showed the similarities with different disease classes even though these questions were considered good ones to separate diseases from each other [Baloh, 1995; Kentala, 1996a]. In addition, the principal component analysis showed that most of the disease classes were extensively overlapping with each other when examining the projection of the two main principal components, hearing and vertigo disorders (Figure 4). These two main principal components explained only 21.7% of the total variance. The mean confusion matrices of the ONE and the 1- and 5-NN OVA having the highest total accuracy in Publication III showed that all disease classes were mixed up with Menière's disease when using the machine learnt knowledge. This happened when using either seven or nine disease classes in the classification. Especially with nine disease classes, sudden deafness (at worst 53.6% with the 5-NN OVA wscat) and central lesion (at worst 44.2% with the 5-NN OVA wscat) cases were confused to Menière's disease. The ONE using pure experts' knowledge mainly misclassified cases as benign recurrent vertigo (at worst 47.1% VNE cases) or benign positional vertigo (at worst 30.0% of BRV cases) but acoustic neurinoma cases were misclassified as sudden deafness (48.1%). The results of Publication V supported also the difficulty of separating the disease classes. With the 5-NN OVA, there occurred tie situations in class voting with 16.2% of the cases. All the tie situations occurred when the classifiers voted the case to be a non-member of the OVA classifiers, in other words, the case's disease class was not separable. With the 5-NN OVO, tie situations occurred only with 2.0% of the cases: The cases of benign positional vertigo, Menière's disease and vestibulopatia were difficult to distinguish from each other. The strong overlap with the disease classes makes the separation of the disease classes challenging even for the machine learning methods.

Due to the difficulty of the otoneurological domain and to the reason that patient can actually have two diseases present at the same time [Kentala *et al.*, 1996], it might be good to check more than one diagnosis suggestions of the ONE to support the diagnosis of a new patient. In the end, the final diagnosis is made by the physician based on the given information on all alternative diseases [Kentala *et al.*, 1996]. The ONE utilizing the equal weighting had 90 cases with the same score and score difference within the first and second diagnosis suggestions and even 12 cases with

the same score and score difference within the first, second and third diagnosis suggestions [II]. The order of suggestions having the same score and score difference was selected randomly and, thus, the first diagnosis suggestion actually could have been any of these two or three suggestions. The total classification accuracies within the three first diagnosis suggestions of the ONE varied from 86.2% (ONE123 experts) to 94.4% (ONE123 wscat) with seven disease classes and from 81.7% (ONE123 wIB4) to 85.0% (ONE123 wscat) with nine disease classes whereas the accuracies with the first diagnosis suggestion varied from 43.3% (ONE1 experts) to 74.6% (ONE1 wscat) with seven disease classes and from 59.1% (ONE1 wIB4) to 62.4% (ONE1 wscat) [III]. Even though in the dissertation was concentrated on the comparison and improvement of the results of the first diagnosis suggestion of the nearest pattern method of the ONE, it is recommended to verify the diagnosis of more than one diagnosis suggestions and their explanations.

A research limitation is recognized. In Publication I, the classification run of the ONE using the original pure experts' knowledge was done with the knowledge base containing descriptions of 15 disease classes. This made the classification harder for the ONE with this knowledge base and partly explains the low classification results with the original pure experts' knowledge. Each knowledge base should have included the same number of disease classes. For the following classification runs, the number of disease classes within the knowledge bases was the same.

As was mentioned in the Chapter 2.2, the graphical user interface, the inference engine and the explanation facility of the ONE form expert system shell that is possible to take in use in new domains. Thus, only the query base and the knowledge base need to be tailored into the new domain. Domain experts are in a key role in tailoring the query and the knowledge bases in use into new domain: Their expertise is needed in defining the questions to be used in data collection and in the class descriptions and later in evaluating the knowledge formed with the machine learning methods. Help of domain experts is needed also in collecting domain data. At first, it is possible to use the ONE to collect data without the knowledge base. When enough data has been collected, it is possible to form the knowledge base with the help of the fitness value formation method and with the attribute weighting methods, for example, with the Scatter method [Juhola and Siermala, 2012]. The expert system shell was preliminary tested with hereditary primary immunodeficiencies to build a Primary ImmunoDeficiency expert system (PIDexpert) [Samarghitean *et al.*, 2008]. The PIDexpert was designed to give a diagnostic picture of these hereditary primary immunodeficiencies based on symptoms, signs, medical history, physical findings

and laboratory tests [Samarghitean and Vihinen, 2008]. In the future, it would be good to test the expert system shell also with data from different domains.

The inference engine and the knowledge base of the ONE is nowadays applied within the Internet-based peer-support program for Menière's disease to verify and assess the diagnosis of person using the peer-support system [Rasku *et al.*, 2015]. The otoneurological questionnaire is used as a basis on the questionnaire of the Menière's disease that includes more questions about the activity limitations, participation restrictions, the International Classification of Functioning, Disability, and Health (ICF) -based problem classification, personality traits (sense of coherence), positive aspects and post-traumatic growth inventory. In the peer-support system, the ONE contains descriptions of 14 different conditions. If the Menière's disease is given by the ONE as the principal diagnosis (the first diagnosis suggestion) and its score is more than 0.43, the person is allowed to attend the peer-support program [Rasku *et al.*, 2015]. Other selection criteria are also given. The peer-support given by the system can be tailored to meet individual needs by the answers given by the user.

One future research plan, in addition to the OVA and the OVO classifiers approach with the ONE, is to divide the current knowledge description of the Menière's disease in the ONE to two or more descriptions. It is acknowledged that the symptoms vary depending of the phase of the disease and, thus, it would be logical to divide it in the early stage Menière's disease and (progressed) Menière's disease to separate the characteristics of them to different descriptions instead of one wide description. For example, as the Menière's disease progresses, vertigo attacks occur more frequently and are more severe than in the beginning of the disease and, also, unilateral auditory symptoms develop in time to bilateral symptoms with almost half of the patients [Chawla and Olshaker, 2006]. The AAO-HNS has proposed to define Menière's disease as "possible Menière's disease", "probable Menière's disease" and "definite Menière's disease" [AAO-HNS, 1995]. However, "certain Menière's disease" needs histological verification of endolymphatic hydrops in the inner ear and, thus, does not help to define the condition clinically [Rasku *et al.*, 2015]. The partition of the current Menière's disease description into two different stage descriptions might also help the recognition of other diseases when description of the Menière's disease would not have such a broad definition.

In the dissertation, different machine learning methods were applied in the classification of the otoneurological data. At the moment, the ONE contains two inference methods, the nearest pattern and the *k*-nearest neighbour methods. Unfortunately, they are not currently applicable at the same time. It would be good to upgrade the ONE as a hybrid decision support system that can utilize different

machine learning methods at the same time to give more support for decision making. In Kononenko *et al.*'s research, physicians felt that the reliability and the comprehensibility of the system was much better when there were utilized more than one machine learning method to support decision making (a multistrategy approach was used) [Kononenko *et al.*, 1998]. The possibilities offered by integrated approaches for multicriteria decision aid [Doumpos and Zopounidis, 2013] should be explored to enhance the classification of the otoneurological diseases and to form an intelligent otoneurological decision support system.

6 REFERENCES

[Aalto, 2005] Aalto P. *Equihear-markkinaselvitys. Kuulon ja huimauksen IT-pohjainen konsepti* [in Finnish]. Finn-Medi Tutkimus, Tampere, Finland, 2005.

[AAO-HNS, 1995] American Academy of Otolaryngology–Head and Neck Surgery (AAO-HNS). Committee on Hearing and Equilibrium guidelines for the diagnosis and evaluation of therapy in Meniere's disease. *Otolaryngology–Head and Neck Surgery* 113(3), 1995, pp. 181–185.

[Aha, 1992] Aha DW. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies* 36(2), 1992, pp. 267–287. https://doi.org/10.1016/0020-7373(92)90018-G

[Aha *et al.*, 1991] Aha DW, Kibler D and Albert MK. Instance-based learning algorithms. *Machine Learning* 6(1), 1991, pp. 37–66.

[Allwein *et al.*, 2000] Allwein EL, Schapire RE and Singer Y. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research* 1, 2000, pp. 113–141.

[Auramo, 1999] Auramo Y. *Construction of an expert system to support otoneurological vertigo diagnosis*. Academic dissertation, Department of Computer Sciences, University of Tampere, Finland, 1999.

[Auramo and Juhola, 1995] Auramo Y and Juhola M. Comparison of inference results of two otoneurological expert systems. *International Journal of Bio-Medical Computing* 39, 1995, pp. 327–335.

[Auramo and Juhola, 1996] Auramo Y and Juhola M. Modifying an expert system construction to pattern recognition solution. *Artificial Intelligence in Medicine* 8, 1996, pp. 15–21.

[Auramo *et al.*, 1993] Auramo Y, Juhola M and Pyykkö I. An expert system for the computer-aided diagnosis of dizziness and vertigo. *Medical Informatics* 18, 1993, pp. 293–305.

[Autio *et al.*, 2007] Autio L, Juhola M and Laurikkala J. On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension. *Computers in Biology and Medicine* 37, 2007, pp. 388–397.

[Baloh, 1995] Baloh RW. Approach to the evaluation of the dizzy patient. *Otolaryngology–Head and Neck Surgery* 112(1), 1995, pp. 3–7.

[Ben-David, 2007] Ben-David A. A lot of randomness is hiding in accuracy. *Engineering Applications of Artificial Intelligence* 20, 2007, pp. 875–885.

[Boháčik and Juhola, 2008] Boháčik J and Juhola M. Fuzzy rule induction and classification applied to otoneurological data. *Journal of Information, Control and Management Systems* 6 (1), 2008, pp. 7–12.

[Cestnik, 1990] Cestnik B. Estimating probabilities: a crucial task in machine learning. In: *Proceedings of the European Conference on Artificial Intelligence (ECAI 1990)*, Stockholm, 1990, pp. 147–149.

[Chawla and Olshaker, 2006] Chawla N and Olshaker JS. Diagnosis and management of dizziness and vertigo. *The Medical Clinics of North America* 90(2), 2006, pp. 291–304.

[Cohen, 1960] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 1960, pp. 37–46.

[Cortes and Vapnik, 1995] Cortes C and Vapnik V. Support-vector networks. *Machine Learning* 20, 1995, pp. 273–297.

[Cover and Hart, 1967] Cover TM and Hart PE. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1), 1967, pp. 21–27.

[De Jong, 1975] De Jong KA. *An analysis of the Behavior of a Class of Genetic Adaptive Systems*. Academic Dissertation, Computer and Communication Sciences Department, University of Michigan, Ann Arbor, USA, 1975. http://hdl.handle.net/2027.42/4507 (Accessed 22nd August 2018)

[Dong *et al.*, 2014] Dong C, Wang Y, Zhang E and Wang N. The methodology of Dynamic Uncertain Causality Graph for intelligent diagnosis of vertigo. *Computer Methods and Programs in Biomedicine* 113(1), 2014, pp. 162–174.

[Doumpos and Zopounidis, 2013] Doumpos M and Zopounidis C. Computational intelligence techniques for multicriteria decision aiding: An overview. In: Doumpos M and Grigoroudis E (eds.), *Multicriteria Decision Aid and Artificial Intelligence: Links, Theory and Applications*. John Wiley & Sons, Ltd, 2013, pp. 2–23.

[Duda *et al.*, 2001] Duda RO, Hart PE and Stork DG. *Pattern Classification.* 2nd Ed., A Wiley-Interscience Publication, John Wiley & Sons, Inc, USA, 2001.

[Exarchos *et al.*, 2016] Exarchos TP, Rigas G, Bibas A, Kikidis D, Nikitas C, Wuyts FL, Ihtijarevic B, Maes L, Cenciarini M, Maurer C, Macdonald N, Bamiou D-E, Luxon L, Prasinos M, Spanoudakis G, Koutsouris DD and Fotiadis DI. Mining balance disorders' data for the development of diagnostic decision support systems. *Computers in Biology and Medicine* 77, 2016, pp. 240–248. https://doi.org/10.1016/j.compbiomed.2016.08.016

[Freund and Schapire, 1997] Freund Y and Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55, 1997, pp. 119–139.

[Friedman, 1996] Friedman JH. *Another approach to polychotomous classification*. Stanford University; 1996.

[Fürnkranz, 2001] Fürnkranz J. Round robin rule learning. In: Brodley CE, Danyluk AP (eds), *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*. Williamstown, MA, Morgan Kauffman, 2001, pp. 146–153.

[Galar *et al.*, 2011] Galar M, Fernández A, Barrenechea E, Bustince H and Herrera F. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition* 44(8), 2011, pp. 1761–1776.

[Gavilán *et al.*, 1990] Gavilán C, Gallego J and Gavilán J. 'Carnisel': an expert system for vestibular diagnosis. *Acta Oto-Laryngologica*, 110(3-4), 1990, pp. 161–167.

[Goldberg, 1989] Goldberg DE. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Boston, MA, USA, 1989.

[Havia, 2004] Havia M. *Menière's Disease Prevalence and Clinical Picture*. Academic dissertation, Department of Otorhinolaryngology, University of Helsinki, Finland. http://ethesis.helsinki.fi/julkaisut/laa/kliin/vk/havia/menieres.pdf (Accessed 19th January 2019).

[Hsu and Lin, 2002] Hsu CW and Lin CJ. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* 13(2), 2002, pp. 415–425.

[Joutsijoki, 2012] Joutsijoki H. *Variations on a Theme: The Classification of Benthic Macroinvertebrates*. Academic dissertation, School of Information Sciences, University of Tampere, Finland, 2012. http://urn.fi/urn:isbn:978-951-44-8953-2 (Accessed 19th January 2019)

[Joutsijoki *et al.*, 2013] Joutsijoki H, Varpa K, Iltanen K and Juhola M. Machine learning approach to an otoneurological classification problem. In: *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2013, pp. 1294–1297. https://doi.org/10.1109/EMBC.2013.6609745

[Juhola, 2008] Juhola M. On machine learning classification of otoneurological data. In: S.K. Andersen *et al.* (eds.), *eHealth Beyond the Horizon – Get IT There.* IOS Press, 2008, pp. 211–216.

[Juhola and Siermala, 2012] Juhola M and Siermala M. A scatter method for data and variable importance evaluation. *Integrated Computer-Aided Engineering* 19(2), 2012, pp. 137–149. https://doi.org/10.3233/ICA-2011-0385

[Juhola *et al.*, 2001] Juhola M, Viikki K, Laurikkala J, Pyykkö I and Kentala E. On classification capability of neural networks: a case study with otoneurological data. In: Patel VL, Rogers R and Haux R (eds.), *Proceedings of the 10th World Congress on Medical Informatics (MEDINFO 2001)*, IOS Press, Amsterdam, 2001, pp. 474–478.

[Kelly and Davis, 1991] Kelly JD and Davis L. A hybrid genetic algorithm for classification. In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI 1991)*, vol. 2, San Francisco, CA, USA, Morgan Kaufmann, 1991, pp. 645–650.

[Kentala, 1996a] Kentala E. Characteristics of six otologic diseases involving vertigo. *American Journal of Otolaryngology* 17(6), 1996, pp. 883–892.

[Kentala, 1996b] Kentala E. *A neurotologic expert system for vertigo and characteristics of six otologic diseases involving vertigo*. Academic dissertation, Department of Otorhinolaryngology, University of Helsinki, Finland, 1996.

[Kentala and Pyykkö, 2000] Kentala E and Pyykkö I. Vestibular schwannoma mimicking Ménière's disease. *Acta Oto-Laryngologica* 120(543), 2000, pp. 17–19.

[Kentala and Pyykkö, 2001] Kentala E and Pyykkö I. Clinical picture of vestibular schwannoma. *Auris Nasus Larynx* 28(1), 2001, pp. 15–22.

[Kentala *et al.*, 1995] Kentala E, Pyykkö I, Auramo Y and Juhola M. Database for vertigo. *Otolaryngology–Head and Neck Surgery* 112(3), 1995, pp. 383–390.

[Kentala *et al.*, 1996] Kentala E, Auramo Y, Pyykkö I and Juhola M. Otoneurological expert system. *Annals of Otology, Rhinology & Laryngology* 105(8), 1996, pp. 654–658.

[Kentala *et al.*, 1998] Kentala E, Auramo Y, Juhola M and Pyykkö I. Comparison between diagnoses of human experts and a neurotologic expert system. *Annals of Otology, Rhinology & Laryngology* 107(2), 1998, pp. 135–140.

[Kentala *et al.*, 1999] Kentala E, Laurikkala J, Pyykkö I and Juhola M. Discovering diagnostic rules from a neurotologic database with genetic algorithms. *Annals of Otology, Rhinology & Laryngology* 108(10), 1999, pp. 948–954.

[Kononenko *et al.*, 1998] Kononenko I, Bratko I and Kukar M. Application of machine learning to medical diagnosis. In: Michalski RS, Bratko I and Kubat M (eds.), *Machine Learning and Data Mining: Methods and Applications*, John Wiley & Sons, Ltd, West Sussex, England, 1998, pp. 389–428.

[Kubat *et al.*, 1998] Kubat M, Bratko I and Michalski RS. A review of machine learning methods. In: Michalski RS, Bratko I and Kubat M (eds.), *Machine Learning and Data Mining: Methods and Applications*, John Wiley & Sons, Ltd, West Sussex, England, 1998, pp. 3–69.

[Landis and Koch, 1977] Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 33(1), 1977, pp. 159–174.

[Laurikkala *et al.*, 2000] Laurikkala J, Kentala E, Juhola M, Pyykkö I and Lammi S. Usefulness of imputation for the analysis of incomplete otoneurological data. *International Journal of Medical Informatics* 58–59, 2000, pp. 235–242.

[Laurikkala *et al.*, 2001] Laurikkala J, Kentala E, Juhola M and Pyykkö I. A novel machine learning program applied to discover otological diagnoses. *Scandinavian Audiology* 30(1), 2001, pp. 100–102.

[Lee *et al.*, 2007] Lee H, Kim E and Park M. A genetic feature weighting scheme for pattern recognition. *Integrated Computer-Aided Engineering* 14(2), 2007, pp. 161–171.

[Liou, 1998] Liou YI. Expert system technology: knowledge acquisition. In: Liebowitz J (ed.), *The Handbook of Applied Expert Systems*. CRC Press LLC, Boca Raton, 1998, pp. 2-1–2-11.

[Metaxiotis and Samouilidis, 2000] Metaxiotis KS and Samouilidis JE. Expert systems in medicine: academic illusion or real power? *Information Management & Computer Security* 8(2), 2000, pp. 75–79. https://doi.org/10.1108/09685220010694017

[Michalewicz, 1992] Michalewicz Z. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, Berlin, Germany, 1992.

[Miettinen and Juhola, 2010] Miettinen K and Juhola M. Classification of otoneurological cases according to Bayesian probabilistic models. *Journal of Medical Systems* 34(2), 2010, pp. 119–130. DOI 10.1007/s10916-008-9223-z

[Mitchell, 1996] Mitchell M. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, Mass, USA, 1996.

[Mitchell, 1997] Mitchell T. *Machine Learning*. McGraw-Hill, New York, USA, 1997.

[Mäkiranta, 2005] Mäkiranta V. *Otoneurologisen tiedonkeruukaavakkeen e-toteutus* [in Finnish]. Engineering thesis, Computer Systems Engineering, Tampere Polytechnic, Finland, 2005.

[Phillips-Wren, 2013] Phillips-Wren G. Intelligent decision support systems. In: Doumpos M and Grigoroudis E (eds.), *Multicriteria Decision Aid and Artificial Intelligence: Links, Theory and Applications*. John Wiley & Sons, Ltd, 2013, pp. 25–44.

[Quinlan, 1993] Quinlan JR. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[Rammazzo *et al.,* 2016] Rammazzo L, Kikidis D, Anwer A, Macdonald N, Kyrodimos E, Maurer C, Wuyts F, Luxon L, Bibas A and Bamiou D-E. EMBalance - validation of a decision support system in the early diagnostic evaluation and management plan formulation of balance disorders in primary care: study protocol of a feasibility randomised controlled trial. *Trials* 17, 2016, pp. 435–444. https://doi.org/10.1186/s13063-016-1568-x

[Rasku *et al.*, 2015] Rasku J, Pyykkö I, Levo H, Kentala E and Manchaiah V. Disease profiling for computerized peer support of Ménière's disease. *JMIR Rehabilitation and Assistive Technologies* 2(2):e9, 2015, pp. 1–12. https://doi.org/10.2196/rehab.4109

[Rifkin and Klautau, 2004] Rifkin R and Klautau A. In defense of one-vs-all classification. *Journal of Machine Learning Research* 5, 2004, pp. 101–141.

[Rutka and Barber, 1986] Rutka JA and Barber HO. Recurrent vestibulopathy: third review. *Journal of Otolaryngology* 15(2), 1986, pp. 105–107.

[Samarghitean and Vihinen, 2008] Samarghitean C and Vihinen M. Medical Expert Systems. *Current Bioinformatics* 3, 2008, pp. 56–65.

[Samarghitean *et al.*, 2008] Samarghitean C, Iltanen K, Varpa K, Helminen M, Juhola M and Vihinen M. PIDexpert - decision support system for primary immunodeficiencies. *Clinical and Experimental Immunology* 154, 2008, p. 162.

[Schmid *et al.*, 1987] Schmid R, Zanocco P, Buizza A, Magenes G, Manfrin M and Mira E. An expert system for the classification of dizziness and vertigo. In: Fox J, Fieschi M and Engelbrecht R (eds.), *Proceedings of the European Conference on Artificial Intelligence in Medicine (AIME 1987)*, Marseilles, vol. 33 of *Lecture Notes in Medical Informatics*, 1987, pp. 45–53.

[Shim *et al.*, 2002] Shim JP, Warkentin M, Courtney JF, Power DJ, Sharda R and Carlsson C. Past, present, and future of decision support technology. *Decision Support Systems* 33(2), 2002, pp. 111–126. https://doi.org/10.1016/S0167-9236(01)00139-7

[Shortliffe, 1976] Shortliffe EH. *Computer-based medical consultations: MYCIN*. Artificial Intelligence series 2, Elsevier, New York, USA, 1976.

[Siermala and Juhola, 2006] Siermala M and Juhola M. Techniques for biased data distributions and variable classification with neural networks applied to otoneurological data. *Computer Methods and Programs in Biomedicine* 81, 2006, pp. 128–136.

[Siermala *et al.*, 2007] Siermala M, Juhola M, Laurikkala J, Iltanen K, Kentala E and Pyykkö I. Evaluation and classification of otoneurological data with new data analysis methods based on machine learning. *Information Sciences* 177(9), 2007, pp. 1963–1976.

[Sintonen, 2001] Sintonen H. The 15D instrument of health-related quality of life: properties and applications. *Annals of Medicine* 33(5), 2001, pp. 328–336.

[Swingler, 1996] Swingler K. *Applying Neural Networks – A Practical Guide*. London Academic Press, 1996.

[Syed, 2014] Syed M. *Attribute weighting in K-nearest neighbor classification*. Master's thesis, School of Information Sciences, University of Tampere, Finland, 2014.

[Tapani, 2008] Tapani MJ. *Observations on modernisation of an otoneurological expert system*. Master's thesis, Department of Computer Sciences, University of Tampere, Finland, 2008.

[Turban, 1993] Turban E. *Decision Support and Expert Systems: Management Support Systems*. Macmillan, New York, USA, 1993.

[Varpa, 2005] Varpa K. *Tietämysjärjestelmien tietämyksen esittäminen ja hankinta sekä huimaustautien päätöstukijärjestelmän ja sen tietämyksen uudistaminen* [in Finnish]. Master's thesis, Department of Computer Sciences, University of Tampere, Finland, 2005.

[Viikki, 2002] Viikki K. *Machine learning on otoneurological data: decision trees for vertigo diseases*. Academic dissertation, Department of Computer Sciences, University of Tampere, Finland, 2002. http://urn.fi/urn:isbn:951-44-5390-5 (Accessed 19th January 2019)

[Viikki and Juhola, 2001] Viikki K and Juhola M. Refining the knowledge base of an otoneurological expert system. In: Crespo J, Maojo V and Martin F (eds.), *Medical Data Analysis*, vol. 2199 of *Lecture Notes in Computer Science*, Springer, Berlin, 2001, pp. 276–281.

[Waterman, 1986] Waterman DA. *A Guide to Expert Systems*. Addison-Wesley, Reading, Massachusetts, 1986.

[Wilson and Martinez, 1997] Wilson RD and Martinez TR. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research* 6, 1997, pp. 1–34.

[Witten *et al.*, 2011] Witten IH, Frank E and Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition, Morgan Kauffmann series in data management systems, Elsevier, MA, USA, 2011.

[Yang and Webb, 2002] Yang Y and Webb GI. A comparative study of discretization methods for Naive-Bayes classifiers. In: *Proceedings of the Pacific Rim Knowledge Acquisition Workshop (PKAW 2002)*, Tokyo, Japan, 2002, pp. 159–173.

7 PERSONAL CONTRIBUTIONS

In the following, the personal contributions of the author of this thesis (below KV) are described for each publication included in this dissertation.

I KV designed and implemented the database and the data transfer method into the ONE and participated in the questionnaire and knowledge base update process with the domain experts. The machine learning method for the fitness value formation was based on the previous work of Kati Iltanen who calculated the fitness values for the research. KV created different knowledge base combinations for the ONE and made the classification runs for these sets. Erna Kentala and Ilmari Pyykkö helped in the collection of the data, update of the questionnaire and answered to the domain questions. Martti Juhola supervised the research. The paper was written by KV.

II The idea of the research was designed collectively by the authors. KV participated in the collection of the new data and made the classification runs of the *k*-NN method and the ONE with different knowledge base and data sets. Kati Iltanen made the runs with the NB classifier. Martti Juhola supervised the research. KV had the main responsibility for writing the paper.

III The idea of the research was designed collectively by the authors. Markku Siermala calculated the Scatter values for the attributes. KV implemented the weight formation methods of Aha's IB4 and its variant IB1w and made the classification runs with the *wk*-NN and the ONE. Kati Iltanen and Martti Juhola supervised the research. KV had the main responsibility for writing the paper.

IV The idea of the research was designed collectively by the authors. KV implemented the genetic algorithm and its evaluation methods the nearest pattern method, the *wk*-NN OVA and the *cwk*-NN into Matlab and made the test runs. Kati Iltanen and Martti Juhola supervised the research. KV had the main responsibility for writing the paper.

V The idea of the research was designed collectively by the authors. KV implemented the *k*-NN variants with the OVO and the OVA classifiers and made the experiments with them. Henry Joutsijoki made the experiments with the SVM variants. Kati Iltanen and Martti Juhola supervised the research. The paper was mainly written by KV. Henry Joutsijoki wrote the description of the SVM and its results.

8 APPENDICES

Appendix I

The otoneurological questionnaire (14 pages)

Otoneurological survey

All your answers are confidential and belong to data protection of case record. It is important for research that you would try to answer to all questions with care.

Personal data

1. **Social security number:** _____

2. **Name:** _____

3. **Address:** _____
_____

4. **Phone number:** _____

5. **Gender:**
Female ☐ Male ☐

6. **Preliminary diagnosis** _____
_____

Please mark the answer that best describes your health. If you don't have these particular symptoms nowadays, please answer the questions based on the situation when you previously had the symptoms.

7. **What symptoms do you have?** *(choose one or several options)*
vertigo ☐ gait difficulties ☐ hearing loss ☐ tinnitus ☐ headache ☐

Onset of symptoms

8. **If you have or have previously had vertigo, hearing loss or tinnitus, with what symptoms did your disease start?** *(choose one or several options)*
vertigo ☐ hearing loss ☐ tinnitus ☐ pressure feeling in the ear ☐ gait difficulties ☐

9. **How old were you when the symptoms began?** _____

If you don't have vertigo, please move on to the question 22.

10. If you have or have previously had vertigo and hearing loss, was there time difference between vertigo and hearing loss? *(choose only one option)*

1 = they started at the same time ☐ | 2 = less than a year ☐ | 3 = 1 - 4 years ☐ | 4 = 5 - 10 years ☐ | 5 = more than 10 years ☐

Vertigo

With vertigo attacks is defined temporary vertigo spells which has little or no symptoms between vertiginous spells.

Constant vertigo means continuous gait difficulties or continuous sensation of vertigo in the head.

11. Do you have these symptoms? *(choose one or several options)*

feeling of rotation ☐ | feeling of floating ☐ | tend to fall ☐ | instability when moving ☐ | blackouts ☐

12. When did the first vertigo symptoms occur? *(choose one option)*

1 = less than a month ☐ | 2 = less than a year ☐ | 3 = 1 - 4 years ☐ | 4 = 5 - 10 years ☐ | 5 = more than 10 years ☐

13. What kind of vertigo you have?

1 = constant ☐ | 2 = spells ☐ | 3 = both ☐

If you have constant vertigo, please move on to the question 19.

14. How often stronger spells of vertigo occur?

1 = less than once a year ☐ | 2 = less than once a month ☐ | 3 = monthly ☐ | 4 = weekly ☐ | 5 = daily ☐

15. If you have vertigo attacks, how long do stronger vertigo attacks last?

1 = 1 - 15 seconds ☐ | 2 = 15 seconds - 5 minutes ☐ | 3 = 5 minutes - 4 hours ☐ | 4 = 4 hours - 24 hours ☐ | 5 = more than a day ☐

16. How severe the vertigo attacks are usually?

1 = mild (does not affect chores at all) ☐ | 2 = weak (affects but can continue working normally) ☐ | 3 = moderate (have to stop working) ☐ | 4 = strong (must rest) ☐ | 5 = very strong (difficulties despite rest) ☐

17. Does vertigo attack include nausea and/or vomiting?

0 = no ☐ | 1 = weak ☐ | 2 = moderate ☐ | 3 = strong ☐ | 4 = very strong vomiting ☐

18. Do you have sudden and strong second or two lasting drop attacks or slips?

0 = no ☐ | 1 = rarely ☐ | 2 = less than once a week ☐ | 3 = weekly ☐ | 4 = daily ☐

19. Does changes in position induce vertigo?

0 = no ☐ 1 = weakly ☐ 2 = moderately ☐ 3 = strongly ☐ 4 = very strongly (falls) ☐

20. Does changes in aerobaric pressure or barotrauma (eg. flying, diving, blowing or sneezing) induce vertigo or balance difficulties?

0 = no ☐ 1 = weakly ☐ 2 = moderately ☐ 3 = strongly ☐ 4 = very strongly (falls) ☐

21. Does physical strain (eg. weight lifting) induce vertigo or balance difficulties?

0 = no ☐ 1 = weakly ☐ 2 = moderately ☐ 3 = strongly ☐ 4 = very strongly (falls) ☐

Mobility

22. Do you have balance or gait difficulties (outside vertigo attacks)?

0 = no ☐ 1 = rarely ☐ 2 = less than once a week ☐ 3 = weekly ☐ 4 = constantly ☐

23. If you have constant unsteadiness (outside vertigo attacks), how strong do you experience those?

0 = no handicap ☐ 1 = weak ☐ 2 = moderate ☐ 3 = strong ☐ 4 = very strong (falls) ☐

24. Walking

0 = I can walk normally ☐ 1 = I can walk with little difficulties ☐ 2 = I can walk with notable difficulties ☐ 3 = I can walk only little ☐ 4 = I am unable to walk ☐

25. Standing up from the chair

0 = normally without hands ☐ 1 = with occasional help by hands ☐ 2 = always with help by hands ☐ 3 = with help from others ☐ 4 = I cannot get up ☐

Hearing loss

26. Has your hearing weakened because of your disease?

0 = no ☐ 1 = in the right ear ☐ 2 = in the left ear ☐ 3 = in both ears ☐

If you don't have hearing loss, please move on to the question 30.

27. If you feel your hearing weakened, how much time has passed from the beginning of hearing loss?

1 = less than a month ☐ 2 = less than a year ☐ 3 = 1 - 4 years ☐ 4 = 5 - 10 years ☐ 5 = more than 10 years ☐

28. Does your hearing fluctuate during the vertigo attacks?

0 = no ☐ 1 = yes ☐

29. How did your hearing loss commence?

1 = suddenly (in few days) ☐ 2 = during few months ☐ 3 = during several years ☐

Tinnitus and hyperacusis

With tinnitus is meant different sounds (eg. hum, pulsating sound etc.) occuring in the ear/head. Hyperacusis means that moderately loud sound induces pain in the ear or is sensed very loud.

30. In which ear you have tinnitus?

0 = I have no tinnitus ☐ 1 = in the right ear ☐ 2 = in the left ear ☐ 3 = bilateral ☐ 4 = tinnitus is in the head ☐

If you don't have tinnitus, please move on to the question 34.

31. When did tinnitus first occur?

1 = less than a month ☐ 2 = less than a year ☐ 3 = 1 - 4 years ☐ 4 = 5 - 10 years ☐ 5 = more than 10 years ☐

32. How much handicap does tinnitus cause for your life?

0 = no handicap ☐ 1 = slight handicap (can do normal chores) ☐ 2 = moderate handicap (affects, but can live normally) ☐ 3 = severe handicap (has to stop chores) ☐ 4 = very severe handicap (constant sleeping disorders) ☐

33. What is the type of tinnitus? *(choose one option)*

1 = hum ☐ 2 = ring ☐ 3 = pulse ☐ 4 = buzz, hiss, stir ☐ 5 = other / several voices ☐

34. Do strong voices hurt (hyperacusis)?

0 = no ☐ 1 = in the right ear ☐ 2 = in the left ear ☐ 3 = in both ears ☐

35. Handicap of hyperacusis

0 = no handicap ☐ 1 = weak ☐ 2 = moderate ☐ 3 = strong ☐ 4 = very strong ☐

36. Do you have pressure feeling in the ear?

0 = no ☐ 1 = in the right ear ☐ 2 = in the left ear ☐ 3 = in both ears ☐

Other symptoms

37. Do you have other symptoms? *(choose one or several options)*

0 = no other symptoms ☐ | 1 = feeling of faint ☐ | 2 = sensation of drunkenness ☐ | 3 = blurring of eyes, growing black ☐ | 4 = feeling of unreality ☐

38. Handicap of above symptoms *(choose one alternative)*

0 = no handicap ☐ | 1 = weak ☐ | 2 = moderate ☐ | 3 = strong ☐ | 4 = very strong ☐

39. Does vertigo, hearing loss or tinnitus cause anxiety, tension or nervousness?

0 = no ☐ | 1 = weakly ☐ | 2 = moderately ☐ | 3 = strongly ☐ | 4 = very strongly ☐

40. Vitality

0 = I feel myself healthy ☐ | 1 = I am somewhat weary or feeble ☐ | 2 = I feel moderately weary or feeble ☐ | 3 = I feel very weary or feeble ☐ | 4 = I feel totally exhausted ☐

Headache

If you do have vertigo spells, please answer next questions based on the headache occurring outside the vertigo spells. Otherwise, answer based on the common situation.

41. Do you have headache and if you do, how long does headache last?

0 = no headache ☐ | 1 = less than 2 hours ☐ | 2 = 2 hours - 24 hours ☐ | 3 = constant headache ☐

If you don't have headache (outside vertigo attacks), please move on to the question 44.

42. How often does headache occur?

1 = less than once a year ☐ | 2 = less than once a month ☐ | 3 = monthly ☐ | 4 = weekly ☐ | 5 = daily ☐

43. Do you have headache during the vertigo attacks?

0 = no ☐ | 1 = slightly ☐ | 2 = moderately ☐ | 3 = much ☐ | 4 = very much ☐

Neurological symptoms

44. Do you suffer from fainting (causing unconsciousness)?

0 = no ☐ | 1 = yes ☐

45. Do you suffer from visual blurring or double vision during vertigo attacks?

0 = no ☐ 1 = yes ☐

46. Do you experience weakness of voice, speech stuttering or entangling (dysarthria) during vertigo attacks?

0 = no ☐ 1 = yes ☐

47. Do you have difficulties in swallowing (cranial nerve palsy)?

0 = no ☐ 1 = yes ☐

48. Do you have touch sensation disturbances in the face (paresthesia in face)?

0 = no ☐ 1 = yes ☐

49. Do you have migraine which is diagnosed by physician?

0 = no ☐ 1 = yes ☐

Alcohol

50. How many restaurant portions of alcohol do you consume in a week?

0 = I don't use alcohol ☐ 1 = less than 4 portions ☐ 2 = 5 - 9 portions ☐ 3 = 10 - 20 portions ☐ 4 = more than 20 portions ☐

Oto- and vestibulotoxic drugs

51. Do you use diuretics or heart medicin?

0 = no ☐ 1 = yes ☐

52. Have you got tubercular or other intravenous medicine (aminoglycocides)?

0 = no ☐ 1 = yes ☐

53. Do you use strong pain killer?

0 = no ☐ 1 = occasionally ☐ 2 = weekly ☐ 3 = daily ☐

54. Have you been treated for malignant tumours?

0 = no ☐ 1 = yes ☐

55. Do you have antidepressive treatment?

0 = no ☐ 1 = yes ☐

56. Do you use other drug treatment for psychiatric disorder?

0 = no ☐ 1 = yes ☐

57. Do you use sleeping pills?

0 = no ☐ 1 = yes ☐

Possible damages of internal ear

58. Have you got any direct trauma to the head or neck, or ear infection, which would have been associated with the beginning of the vertigo symptoms? (symptoms occured within 6 months of the event)

0 = no ☐ 1 = yes ☐

59. Have you had any brain concussion with unconsciousness lasting less than 2 hours?

0 = no ..... ☐

1 = yes, ... ☐

in what year? _____

60. Have you had any head injury causing unconsciousness lasting more than 2 hours?

0 = no ..... ☐

1 = yes, ... ☐

in what year? _____

61. Have you had any whiplash injury in the neck?

0 = no ..... ☐

1 = yes, ... ☐

in what year? _____

62. Have you had prolonged (over three months) ear discharge / running ear caused by inflammatio?

0 = no ☐ 1 = yes ☐

63. Have you had any direct trauma to the ear, acute noise injury, bleeding from the ear which would have caused hearing loss or tinnitus?

0 = no ..... ☐

1 = yes, ... ☐

in what year? _____

64. Have you been exposed at work to loud noise (noise level exceeding 85 dB) more than 5 years?

0 = no ☐ 1 = yes ☐

Ear surgery

65. Has your ear(s) been operated?

0 = no ☐ 1 = I don't know ☐ 2 = yes ☐

If you haven't been in ear operations, please move on to the question 68.

66. Which ear has been operated?

1 = the right ear ☐ 2 = the left ear ☐ 3 = both ears ☐ 4 = I don't know ☐

If you know what has been operated, please answer to following question. If you don't know, please move on to the question 68.

67. Have you been at the ear surgery because of the vertigo?

0 = no ..... ☐

1 = yes, ... ☐

in what surgery? _____

in what year? _____

Other diseases

68. Do you have coronary heart disease?

0 = no ☐ 1 = yes ☐

69. Do you have hypertension?

0 = no ☐ 1 = yes ☐

70. Do you have arteriosclerosis?

0 = no ☐ 1 = yes ☐

71. Do you have any symptoms of cerebral or brain stem ischemia?

0 = no ☐ 1 = yes ☐

72. Do you have kidney insufficiency/renal failure ?

0 = no ☐ 1 = yes ☐

73. Do you have diabetes mellitus?

0 = no ☐ 1 = yes ☐

74. Do you have thyroid gland over- or underproduction?

0 = no ☐ 1 = yes ☐

75. Have you ever suffered from meningitis or sequelae of mumps?

0 = no ..... ☐

1 = yes, ... ☐

what? _____

in what year? _____

Family history

76. Does your father or mother have had vertigo or early onset of hearing loss before age 65?

0 = no ☐ 1 = I don't know ☐ 2 = yes ☐

77. Do your siblings have vertigo or early onset of hearing loss before age 65?

0 = no ☐ 1 = I don't know ☐ 2 = yes ☐

78. Do your children have hearing loss?

0 = no ☐ 1 = I don't know ☐ 2 = yes ☐

79. If yes, do you know the reason for vertigo or hearing loss?

1 = I don't know ..... ☐

2 = yes, ... ☐

what? _____

80. How many siblings do you have?

0 = none ☐ 1 = one ☐ 2 = two ☐ 3 = three ☐ 4 = more than three ☐

81. From which county or city are your mother's mother from?

_____

QUALITY OF LIFE QUESTIONNAIRE (15D©)

Please read through all the alternative responses to each question before placing a cross (x) against the alternative which **best describes your present status**. Continue through all 15 questions in this manner, giving only **one** answer to each.

Question 1. Mobility

☐ 1. I am able to walk normally (without difficulty) indoors, outdoors and on stairs.

☐ 2. I am able to walk without difficulty indoors, but outdoors and/or on stairs I have slight difficulties.

☐ 3. I am able to walk without help indoors (with or without an appliance), but outdoors and/or on stairs only with considerable difficulty or with help from others.

☐ 4. I am able to walk indoors only with help from others.

☐ 5. I am completely bed-ridden and unable to move about.

Question 2. Vision

☐ 1. I see normally, i.e. I can read newspapers and TV text without difficulty (with or without glasses).

☐ 2. I can read papers and/or TV text with slight difficulty (with or without glasses).

☐ 3. I can read papers and/or TV text with considerable difficulty (with or without glasses).

☐ 4. I cannot read papers or TV text either with glasses or without, but I can see enough to walk about without guidance.

☐ 5. I cannot see enough to walk about without a guide, i.e. I am almost or completely blind.

Question 3. Hearing

☐ 1. I can hear normally, i.e. normal speech (with or without a hearing aid).

☐ 2. I hear normal speech with a little difficulty.

☐ 3. I hear normal speech with considerable difficulty; in conversation I need voices to be louder than normal.

☐ 4. I hear even loud voices poorly; I am almost deaf.

☐ 5. I am completely deaf.

Question 4. Breathing

☐ 1. I am able to breathe normally, i.e. with no shortness of breath or other breathing difficulty.

☐ 2. I have shortness of breath during heavy work or sports, or when walking briskly on flat ground or slightly uphill.

☐ 3. I have shortness of breath when walking on flat ground at the same speed as others my age.

☐ 4. I get shortness of breath even after light activity, e.g. washing or dressing myself.

☐ 5. I have breathing difficulties almost all the time, even when resting.

Question 5. Sleeping

☐ 1. I am able to sleep normally, i.e. I have no problems with sleeping.

☐ 2. I have slight problems with sleeping, e.g. difficulty in falling asleep, or sometimes waking at night.

☐ 3. I have moderate problems with sleeping, e.g. disturbed sleep, or feeling I have not slept enough.

☐ 4. I have great problems with sleeping, e.g. having to use sleeping pills often or routinely, or usually waking at night and/or too early in the morning.

☐ 5. I suffer severe sleeplessness, e.g. sleep is almost impossible even with full use of sleeping pills, or staying awake most of the night.

Question 6. Eating

☐ 1. I am able to eat normally, i.e. with no help from others.

☐ 2. I am able to eat by myself with minor difficulty (e.g. slowly, clumsily, shakily, or with special appliances).

☐ 3. I need some help from another person in eating.

☐ 4. I am unable to eat by myself at all, so I must be fed by another person.

☐ 5. I am unable to eat at all, so I am fed either by tube or intravenously.

Question 7. Speech

☐ 1. I am able to speak normally, i.e. clearly, audibly and fluently.

☐ 2. I have slight speech difficulties, e.g. occasional fumbling for words, mumbling, or changes of pitch.

☐ 3. I can make myself understood, but my speech is e.g. disjointed, faltering, stuttering or stammering.

☐ 4. Most people have great difficulty understanding my speech.

☐ 5. I can only make myself understood by gestures

Question 8. Excretion

☐ 1. My bladder and bowel work normally and without problems.

☐ 2. I have slight problems with my bladder and/or bowel function, e.g. difficulties with urination, or loose or hard bowels.

☐ 3. I have marked problems with my bladder and/or bowel function, e.g. occasional `accidents', or severe constipation or diarrhea.

☐ 4. I have serious problems with my bladder and/or bowel function, e.g. routine `accidents', or need of catheterization or enemas.

☐ 5. I have no control over my bladder and/or bowel function.

Question 9. Usual activities

☐ 1. I am able to perform my usual activities (e.g. employment, studying, housework, free-time activities) without difficulty.

☐ 2. I am able to perform my usual activities slightly less effectively or with minor difficulty.

☐ 3. I am able to perform my usual activities much less effectively, with considerable difficulty, or not completely.

☐ 4. I can only manage a small proportion of my previously usual activities.

☐ 5. I am unable to manage any of my previously usual activities.

Question 10. Mental function

☐ 1. I am able to think clearly and logically, and my memory functions well

☐ 2. I have slight difficulties in thinking clearly and logically, or my memory sometimes fails me.

☐ 3. I have marked difficulties in thinking clearly and logically, or my memory is somewhat impaired.

☐ 4. I have great difficulties in thinking clearly and logically, or my memory is seriously impaired.

☐ 5. I am permanently confused and disoriented in place and time.

Question 11. Discomfort and symptoms

☐ 1. I have no physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.

☐ 2. I have mild physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.

☐ 3. I have marked physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.

☐ 4. I have severe physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.

☐ 5. I have unbearable physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.

Question 12. Depression

☐ 1. I do not feel at all sad, melancholic or depressed.

☐ 2. I feel slightly sad, melancholic or depressed.

☐ 3. I feel moderately sad, melancholic or depressed.

☐ 4. I feel very sad, melancholic or depressed.

☐ 5. I feel extremely sad, melancholic or depressed.

Question 13. Distress

☐ 1. I do not feel at all anxious, stressed or nervous.

☐ 2. I feel slightly anxious, stressed or nervous.

☐ 3. I feel moderately anxious, stressed or nervous.

☐ 4. I feel very anxious, stressed or nervous.

☐ 5. I feel extremely anxious, stressed or nervous.

Question 14. Vitality

- [ ] 1. I feel healthy and energetic
- [ ] 2. I feel slightly weary, tired or feeble.
- [ ] 3. I feel moderately weary, tired or feeble.
- [ ] 4. I feel very weary, tired or feeble, almost exhausted.
- [ ] 5. I feel extremely weary, tired or feeble, totally exhausted.

Question 15. Sexual activity

- [ ] 1. My state of health has no adverse effect on my sexual activity.
- [ ] 2. My state of health has a slight effect on my sexual activity.
- [ ] 3. My state of health has a considerable effect on my sexual activity.
- [ ] 4. My state of health makes sexual activity almost impossible.
- [ ] 5. My state of health makes sexual activity impossible.

FOOTNOTES (please refer to questionnaire questions with numbers):

If you have been at posturography or audiometry, could you please add the copy with this questionnaire, thank you.

Appendix II

Utilized attributes in the HUCH data. Class-wise minimum and maximum values and percent of missing values.

* Number of corresponding question in the otoneurological questionnaire (App. I).

*	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
7a.	SYM_VERT	Min	0	1	1	0	1	1	1	0	1	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7b.	SYM_MOVDIF	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.0	1.7	1.1	2.1	0.0	1.3	0.0	0.0	0.0	1.0
7c.	SYM_HEARLOSS	Min	0	0	0	1	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.0	0.6	0.0	2.1	1.4	0.0	0.0	0.0	0.0	0.3
7d.	SYM_TINNITUS	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.0	0.6	0.0	0.0	1.4	0.6	0.0	0.0	0.0	0.3
7e.	SYM_HEADACHE	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	4.6	0.6	6.6	4.3	1.4	1.9	0.0	0.0	0.0	3.5
9.	SYM_AGE	Min	13	10	11	24	4	8	12	9	11	4
		Max	73	80	79	82	72	73	59	76	80	82
		Missing %	5.3	0.6	1.4	19.1	1.4	1.3	0.0	5.5	8.3	2.9
11a.	ROTATION	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	3.1	11.0	8.9	14.9	8.2	8.3	5.0	12.7	8.3	8.7
11b.	FLOATING	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	3.1	34.1	8.6	12.8	6.8	31.2	5.0	18.2	8.3	16.1
12.	AGE_SYMPTOMS	Min	0	1	1	0	1	1	1	0	1	0
		Max	4	4	4	4	4	4	4	4	4	4
		Missing %	0.8	1.2	1.1	0.0	4.1	2.5	0.0	1.8	4.2	1.6
14.	ATT_OFTEN	Min	0	1	1	0	1	1	1	0	1	0
		Max	5	5	5	5	5	5	5	5	5	5
		Missing %	20.6	3.5	3.1	8.5	4.1	0.6	0.0	3.6	8.3	5.4
15.	ATT_LAST	Min	0	1	0	0	1	1	1	0	1	0
		Max	5	5	5	5	5	5	5	5	5	5
		Missing %	22.1	0.0	4.9	8.5	5.5	5.7	10.0	12.7	29.2	7.7

	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
16.	ATT_INTE	Min	0	2	0	0	2	2	3	0	2	0
		Max	5	5	5	5	5	5	5	5	5	5
		Missing %	3.1	2.9	1.1	4.3	4.1	5.1	5.0	3.6	16.7	3.2
17.	NAUSEA	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	3	3	3	3	3	3	3	3	3
		Missing %	0.8	0.6	1.4	2.1	4.1	1.3	0.0	3.6	8.3	1.7
18.	SLIPSFALLS	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	4	4	3	4	4	3	4	4	4
		Missing %	0.0	4.0	1.7	0.0	1.4	3.8	10.0	3.6	8.3	2.5
19.	PROV_POSIT	Min	0	0	0	0	0	0	0	0	0	0
		Max	4	4	4	4	4	4	4	4	4	4
		Missing %	0.8	4.6	3.1	4.3	5.5	5.7	5.0	7.3	0.0	3.9
20.	PROV_PRESS	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	4	4	4	4	4	3	4	4	4
		Missing %	1.5	4.6	1.7	6.4	2.7	6.4	0.0	7.3	4.2	3.5
21.	PROV_PHYSIC	Min	0	0	0	0	0	0	0	0	0	0
		Max	4	4	4	4	4	4	3	4	4	4
		Missing %	1.5	3.5	4.0	2.1	2.7	8.9	0.0	5.5	4.2	4.2
22.	UNSTEADINESS	Min	0	0	0	0	0	0	0	0	0	0
		Max	4	3	4	1	3	4	3	4	4	4
		Missing %	6.9	3.5	8.6	46.8	5.5	4.5	0.0	0.0	0.0	7.6
26.	HL_SIDE	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	3	3	3	3	3	3	3	3	3
		Missing %	2.3	6.4	4.9	6.4	12.3	7.0	0.0	0.0	0.0	5.2
27.	AGE_HL_SYM	Min	0	0	0	1	0	0	0	0	0	0
		Max	4	4	4	4	4	4	4	4	4	4
		Missing %	6.1	8.7	2.0	4.3	5.5	7.0	0.0	5.5	0.0	4.9
28.	HL_FLUCT	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	4.6	9.8	12.0	19.1	12.3	7.0	0.0	1.8	0.0	9.2
29.	HL_TYPE	Min	0	0	0	1	0	0	0	0	0	0
		Max	2	2	2	1	2	2	2	2	2	2
		Missing %	87.8	54.3	58.6	14.9	34.2	62.4	0.0	1.8	4.2	53.0
30.	TINNI_LOC	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	3	3	3	3	3	3	3	3	3
		Missing %	1.5	7.5	2.0	2.1	5.5	8.9	5.0	1.8	8.3	4.4
31.	AGE_TIN_SYM	Min	0	0	0	0	0	0	0	0	0	0
		Max	4	4	4	4	4	4	4	4	4	4
		Missing %	6.9	10.4	2.6	8.5	2.7	12.7	5.0	3.6	16.7	6.7
32.	TINNITUS	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	3	3	3	3	3	1	3	3	3
		Missing %	0.0	2.9	1.1	0.0	0.0	6.4	0.0	0.0	0.0	1.8
37e.	LIGHTHEAD	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.0	1.2	0.3	2.1	0.0	1.3	0.0	0.0	0.0	0.6

	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
38.	OTHER_HARM	Min	0	0	0	0	0	0	0	0	0	0
		Max	4	4	4	3	4	4	4	4	4	4
		Missing %	1.5	5.8	3.7	8.5	6.8	5.1	0.0	5.5	0.0	4.4
39.	ANXIETY	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	2	3	3	2	3	2	3	2	3
		Missing %	6.9	6.4	6.9	6.4	6.8	3.8	0.0	0.0	4.2	5.7
42.	HA_OCCUR	Min	0	0	0	0	0	0	0	0	0	0
		Max	5	5	5	5	5	5	5	5	5	5
		Missing %	13.0	6.4	17.7	8.5	13.7	6.4	0.0	0.0	0.0	11.1
43.	NEUR_HA	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	3	3	3	3	3	3	3	3	3
		Missing %	2.3	4.0	4.3	6.4	8.2	3.8	0.0	3.6	4.2	4.2
44.	NEUR_SYNCOPE	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	0	1	1	1	1	1	1
		Missing %	0.8	6.4	9.7	4.3	4.1	8.3	0.0	1.8	4.2	6.4
45.	NEUR_VISUAL	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.8	13.3	7.7	51.1	8.2	7.0	5.0	7.3	8.3	9.6
46.	NEUR_DYSA	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.0	10.4	6.6	8.5	6.8	10.2	0.0	3.6	8.3	6.8
47.	NEUR_CNP	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	1.5	11.6	4.3	51.1	8.2	7.6	5.0	5.5	16.7	8.4
48.	NEUR_PARES	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.0	11.0	4.6	8.5	9.6	7.6	5.0	5.5	12.5	6.3
50.	ALCOHOL	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	3	3	3	3	3	2	3	3	3
		Missing %	68.7	0.6	3.4	2.1	4.1	3.2	0.0	0.0	0.0	10.9
51.	OTO_DIUR	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	1.5	0.6	2.9	2.1	2.7	3.8	0.0	0.0	0.0	2.1
52.	OTO_AMINO	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	0	1	0	0	0	0	0	0	1
		Missing %	1.5	0.6	4.9	2.1	4.1	3.2	0.0	0.0	0.0	2.8
53.	OTO_NSAD	Min	0	0	0	0	0	0	0	0	0	0
		Max	3	3	3	1	3	3	2	3	1	3
		Missing %	1.5	2.3	6.6	44.7	8.2	4.5	0.0	3.6	12.5	6.6
54.	OTO_CYTO	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	1	1	0	0	0	0	1	0	1
		Missing %	0.8	1.2	4.0	6.4	5.5	3.2	0.0	0.0	0.0	2.8
55.	TRICY_ANTID	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	0	1	0	0	1	1	1
		Missing %	1.5	3.5	6.0	2.1	9.6	6.4	0.0	0.0	0.0	4.6

	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
56.	KLORPROMAZIN	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	0	1	1	0	0	0	1
		Missing %	1.5	4.0	6.3	2.1	9.6	6.4	0.0	0.0	0.0	4.8
57.	BARBITURATE	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	0	0	0	0	1
		Missing %	1.5	3.5	6.6	2.1	9.6	6.4	0.0	0.0	0.0	4.8
58.	HEAD_TRAUMA	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	1	1	0	1	1	0	0	1	1
		Missing %	3.1	4.6	3.4	4.3	1.4	5.1	0.0	0.0	0.0	3.4
59.	CONCUSSION	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	1	1	1	1	1	0	0	0	1
		Missing %	0.8	4.0	4.0	6.4	0.0	5.1	0.0	0.0	0.0	3.2
60.	CONTUSION	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	0	0	0	1	0	0	1	1	1
		Missing %	0.8	4.6	3.4	6.4	0.0	5.1	0.0	0.0	0.0	3.1
61.	WHIP_INJ	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	1	1	0	1	1	0	0	0	1
		Missing %	0.8	4.0	3.4	6.4	1.4	5.1	0.0	0.0	0.0	3.1
62.	EAR_HIST	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	0	1	1	1
		Missing %	4.6	6.4	8.9	4.3	11.0	7.0	0.0	0.0	0.0	6.7
63.	EARTRAUMA	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	1	1	0	1	0	1	0	0	1
		Missing %	4.6	5.2	7.4	4.3	5.5	5.7	0.0	0.0	0.0	5.4
64.	NOISEEXP	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	4.6	4.6	7.4	4.3	5.5	5.7	0.0	0.0	0.0	5.3
65.	EO_DONE	Min	0	0	0	0	0	0	0	0	0	0
		Max	2	2	2	2	2	2	2	2	2	2
		Missing %	77.1	3.5	7.1	2.1	4.1	1.9	0.0	0.0	0.0	13.5
68.	HEART_ISH	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	0	1	1	1	1	1
		Missing %	1.5	8.1	5.4	6.4	5.5	8.9	0.0	1.8	4.2	5.6
69.	HYPERTEN	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	1	1	1	1
		Missing %	0.8	5.2	4.3	8.5	4.1	7.0	5.0	0.0	0.0	4.3
70.	ANTER_SCL	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	0	0	1	0	1	1	1
		Missing %	2.3	6.4	4.6	46.8	4.1	8.9	0.0	0.0	0.0	6.7
71.	BRAIN_ISH	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	0	1	1	0	1	1	1
		Missing %	0.8	10.4	8.3	51.1	8.2	10.8	0.0	18.2	4.2	10.3
72.	KIDNEY_INS	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	0	1	0	0	1	1	1	0	1
		Missing %	0.8	6.9	5.1	10.6	4.1	6.4	0.0	0.0	0.0	4.8

	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
73.	DIABETES	Min	0	0	0	0	0	0	0	0	0	0
		Max	1	1	1	1	1	1	0	0	0	1
		Missing %	0.8	5.8	4.3	10.6	4.1	6.4	0.0	0.0	0.0	4.3
74.	THYROID	Min	0	0	0	0	0	0	0	0	0	0
		Max	0	1	1	0	0	1	1	1	1	1
		Missing %	1.5	7.5	5.4	48.9	4.1	7.0	0.0	0.0	0.0	6.9

Clinical test results

	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
	SP_NYST	Min	0	0	0	0	0	0	0	0	0	0
	spontanic	Max	1	1	1	1	1	1	1	1	0	1
	nystagmus	Missing %	19.8	42.8	29.4	34.0	30.1	29.9	25.0	5.5	20.8	29.2
	HEAD_SHAK	Min	0	0	0	0	0	0	0	0	0	0
	head shaking	Max	1	1	1	1	1	1	1	1	1	1
	nystagmus	Missing %	20.6	43.9	31.1	36.2	32.9	26.1	25.0	7.3	20.8	29.9
	FING_NOSE	Min	0	0	0	0	0	0	0	0	0	0
	abnormal	Max	1	0	1	1	1	0	0	0	1	1
	finger-nose test	Missing %	19.8	44.5	33.7	38.3	38.4	20.4	25.0	20.0	33.3	31.4
	DIADOCHOKIN	Min	0	0	0	0	0	0	0	0	0	0
	abnormal	Max	1	0	0	0	0	0	0	1	0	1
	diadochokinesis	Missing %	24.4	46.8	40.6	38.3	41.1	22.3	35.0	27.3	37.5	35.8
	POST_OPEN	Min	0.8	0.7	0.5	0.9	0.6	0.7	1.1	0.8	1	0.5
	posturography	Max	10	3.2	4.2	2.9	4.25	4.1	2.8	6.7	6	10.0
	eyes open	Missing %	24.4	28.9	32.6	36.2	23.3	33.1	55.0	50.9	54.2	32.4
	POST_CLOSE	Min	1	0.7	0.75	1.1	1	0.9	1	0.9	1.2	0.7
	posturography	Max	10.1	10	10	4.2	5.2	7.2	4.9	9.7	7.2	10.1
	eyes closed	Missing %	23.7	29.5	32.6	38.3	23.3	33.1	55.0	52.7	54.2	32.6
	CAL_SP_NYST	Min	0	0	0	0	0	0	0	0	0	0
	ENG spontanic	Max	6	7	10	8	9	10	3	9	5	10
	nystagmus	Missing %	19.8	10.4	12.9	14.9	13.7	8.3	10.0	12.7	8.3	12.6
	CAL_ASYM	Min	2	0	0	0	0	0	2	0	0	0
	ENG caloric	Max	100	98	100	100	100	100	100	100	32	100
	asymmetry	Missing %	13.0	10.4	10.9	10.6	6.8	3.8	15.0	14.5	4.2	9.8
	CAL_44R	Min	0	0	0	0	0	0	0	0	8	0
	ENG response with	Max	50	50	50	50	44	50	38	50	42	50
	44°C right	Missing %	67.2	7.5	9.4	8.5	4.1	2.5	15.0	16.4	8.3	15.4
	CAL_44L	Min	0	0	0	0	0	0	4	0	6	0
	ENG response with	Max	50	50	50	50	50	50	32	50	50	50
	44°C left	Missing %	69.5	7.5	9.4	8.5	4.1	4.5	20.0	14.5	8.3	16.0
	AUD_500R	Min	0	0	0	0	0	0	0	0	0	0
	audiometry at	Max	120	80	120	100	80	100	20	100	35	120
	500 Hz right	Missing %	2.3	3.5	3.7	6.4	5.5	0.6	5.0	1.8	4.2	3.2
	AUD_1000R	Min	0	0	0	0	0	0	0	0	0	0
	audiometry at	Max	120	100	120	95	95	100	20	100	45	120
	1000 Hz right	Missing %	3.1	3.5	3.7	6.4	5.5	0.6	5.0	1.8	4.2	3.3

	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
	AUD_2000R	Min	0	0	0	0	0	0	0	0	0	0
	audiometry at	Max	120	100	120	95	100	100	45	100	55	120
	2000 Hz right	Missing %	2.3	3.5	3.7	6.4	5.5	0.6	5.0	1.8	4.2	3.2
	AUD_4000R	Min	0	0	0	0	0	0	0	0	0	0
	audiometry at	Max	90	100	100	100	100	100	80	100	65	100
	4000 Hz right	Missing %	84.7	17.9	28.0	19.1	20.5	23.6	5.0	1.8	4.2	29.5
	AUD_8000R	Min	0	0	0	0	0	0	0	0	5	0
	audiometry at	Max	95	100	100	100	100	100	75	100	85	100
	8000 Hz right	Missing %	84.7	17.9	28.0	19.1	20.5	23.6	5.0	1.8	4.2	29.5
	AUD_500L	Min	0	0	0	0	0	0	0	0	0	0
	audiometry at	Max	120	110	100	110	100	90	25	100	85	120
	500 Hz left	Missing %	3.8	4.0	4.3	2.1	5.5	0.6	5.0	1.8	4.2	3.5
	AUD_1000L	Min	0	0	0	0	0	0	0	0	0	0
	audiometry at	Max	120	120	100	120	100	95	30	100	70	120
	1000 Hz left	Missing %	2.3	3.5	4.3	2.1	5.5	0.6	5.0	1.8	4.2	3.2
	AUD_2000L	Min	0	0	0	0	0	0	0	0	0	0
	audiometry at	Max	120	120	105	120	100	80	40	100	80	120
	2000 Hz left	Missing %	2.3	3.5	4.3	2.1	5.5	0.6	5.0	1.8	4.2	3.2
	AUD_4000L	Min	5	0	0	5	0	0	0	0	0	0
	audiometry at	Max	100	100	100	100	100	75	70	100	100	100
	4000 Hz left	Missing %	84.7	18.5	28.3	14.9	20.5	23.6	5.0	1.8	4.2	29.5
	AUD_8000L	Min	0	0	0	10	0	0	0	0	0	0
	audiometry at	Max	100	100	100	100	100	85	75	100	100	100
	8000 Hz left	Missing %	84.7	17.9	28.3	14.9	21.9	23.6	5.0	1.8	4.2	29.5

Derived attributes

	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
	CNS_SYMPTOMS	Min	0	0	0	0	0	0	0	0	0	0
	neurological	Max	1	1	1	1	1	1	1	1	1	1
	symptoms	Missing %	1.5	2.9	4.9	6.4	4.1	2.5	0.0	3.6	0.0	3.5
	OTO_HABIT_DRUGS	Min	0	0	0	0	0	0	0	0	0	0
	use of oto- and vesti-	Max	1	1	1	1	1	1	1	1	1	1
	bulotoxic drugs	Missing %	3.8	1.2	4.6	2.1	4.1	1.3	0.0	1.8	4.2	3.0
	INJURY	Min	0	0	0	0	1	0	0	0	0	0
	head or ear trauma,	Max	0	1	1	1	1	1	1	1	1	1
	noise injury	Missing %	0.0	1.2	5.4	0.0	0.0	1.3	0.0	0.0	0.0	2.2
	EAR_ILLNESS	Min	0	0	0	0	0	0	0	0	0	0
	Ear infections	Max	1	1	1	1	1	1	0	1	1	1
		Missing %	4.6	3.5	6.0	2.1	5.5	1.3	0.0	0.0	0.0	3.9
	GEN_ILLNESS	Min	0	0	0	0	0	0	0	0	0	0
	disease provoking	Max	1	1	1	1	1	1	1	1	1	1
	vertigo	Missing %	0.0	4.0	3.4	12.8	4.1	0.6	0.0	9.1	0.0	3.3
	BILATERAL	Min	0	0	0	0	0	0	0	0	0	0
	bilateral hearing loss	Max	1	1	1	1	1	1	1	1	1	1
	/tinnitus/hyperacusis / pressure feeling in the ear	Missing %	0.8	3.5	0.0	0.0	4.1	0.6	5.0	1.8	0.0	1.3

	Attribute name		ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Total
	UNILATERAL	Min	0	0	0	0	0	0	0	0	0	0
	unilateral hearing loss	Max	1	1	1	1	1	1	1	1	1	1
	/tinnitus/hyperacusis /	Missing %	0.8	3.5	0.0	0.0	4.1	0.6	5.0	1.8	0.0	1.3
	pressure feeling in the ear											
	BILAT_KA	Min	0	0	0	0	0	0	0	0	0	0
	bilateral hearing loss	Max	1	1	1	1	1	1	1	1	1	1
		Missing %	3.8	8.7	8.0	10.6	17.8	7.0	5.0	1.8	4.2	7.8
	UNILAT_KA	Min	0	0	0	0	0	0	0	0	0	0
	unilateral hearing loss	Max	1	1	1	1	1	1	1	1	1	1
		Missing %	3.8	8.7	8.0	10.6	17.8	7.0	5.0	1.8	4.2	7.8
	LAT_KA	Min	0	0	0	0	0	0	0	0	0	0
	bi- or unilateral	Max	1	1	1	1	1	1	1	1	1	1
	hearing loss	Missing %	1.5	8.7	4.3	4.3	16.4	7.0	5.0	1.8	0.0	5.7
	NONLAT_KA	Min	0	0	0	0	0	0	0	0	0	0
	normal hearing	Max	1	1	1	1	1	1	1	1	1	1
		Missing %	1.5	8.7	4.3	4.3	16.4	7.0	5.0	1.8	0.0	5.7
	VERTIGO	Min	0	1	0	0	0	1	1	0	1	0
	true vertigo	Max	1	1	1	1	1	1	1	1	1	1
		Missing %	3.1	12.1	7.4	12.8	6.8	9.6	0.0	12.7	8.3	8.3
	HL_MEN	Min	0	0	0	0	0	0	0	0	0	0
	hearing side difference	Max	1	1	1	1	1	1	1	1	1	1
	15 dB in 500 Hz, 1000 Hz	Missing %	2.3	3.5	4.6	6.4	5.5	0.6	5.0	1.8	4.2	3.5
	or 2000 Hz											
	HL_ANE	Min	0	0	0	0	0	0	0	0	0	0
	hearing side difference	Max	1	1	1	1	1	1	1	1	1	1
	15 dB in any frequency	Missing %	2.3	3.5	4.6	6.4	5.5	0.6	5.0	1.8	4.2	3.5
	TRAUMA	Min	0	0	0	0	1	0	0	0	0	0
	serious trauma of	Max	0	1	1	1	1	1	0	1	1	1
	head	Missing %	3.1	4.0	4.3	6.4	0.0	5.1	0.0	0.0	0.0	3.6
	AGES_SAME	Min	0	0	0	0	0	0	0	0	0	0
	vertigo, hearing loss	Max	1	1	1	1	1	1	1	1	1	1
	and tinnitus started	Missing %	5.3	5.8	2.9	8.5	5.5	9.6	0.0	5.5	4.2	5.2
	at the same time											

PUBLICATIONS

Publication I Varpa K, Iltanen K, Juhola M, Kentala E and Pyykkö I. Refinement of the otoneurological decision support system and its knowledge acquisition process. In: Engelbrecht R and Hasman A (eds.), *European Notes in Medical Informatics: Ubiquity: Technologies for Better Health in Aging Societies, vol. II, no. 2, 2006. Proceedings of the 20th International Congress of the European Federation for Medical Informatics (MIE 2006)*, Maastricht, Netherlands, 2006, pp. 197–202.

Publication II Varpa K, Iltanen K and Juhola M. Machine learning method for knowledge discovery experimented with otoneurological data. *Computer Methods and Programs in Biomedicine* 91(2), 2008, pp. 154–164. https://doi.org/10.1016/j.cmpb.2008.03.003

Publication III Varpa K, Iltanen K, Siermala M and Juhola M. Attribute weighting with Scatter and instance-based learning methods evaluated with otoneurological data. *International Journal of Data Science* 2(3), 2017, pp. 173–204. https://doi.org/10.1504/IJDS.2017.10007392

Publication IV Varpa K, Iltanen K and Juhola M. Genetic algorithm based approach in attribute weighting for a medical data set. *Journal of Computational Medicine* 2014, 2014, pp. 1–11. https://doi.org/10.1155/2014/526801

Publication V Varpa K, Joutsijoki H, Iltanen K and Juhola M. Applying one-vs-one and one-vs-all classifiers in *k*-nearest neighbour method and support vector machines to an otoneurological multi-class problem. In: Moen A *et al.* (eds.), *Studies in Health Technology and Informatics vol. 169, 2011: User Centred Networked Health Care – Proceedings of 23rd International Conference of the European Federation for Medical Informatics (MIE 2011)*, Oslo, Norway, IOS Press, 2011, pp. 579–583. https://doi.org/10.3233/978-1-60750-806-9-579

PUBLICATION I

Refinement of the Otoneurological Decision Support System and its Knowledge Acquisition Process

Kirsi Varpa, Kati Iltanen, Martti Juhola, Erna Kentala and Ilmari Pyykkö

In: Engelbrecht R and Hasman A (eds.), *European Notes in Medical Informatics: Ubiquity: Technologies for Better Health in Aging Societies, vol. II, no. 2, 2006. Proceedings of the 20th International Congress of the European Federation for Medical Informatics (MIE 2006)*, Maastricht, Netherlands, 2006, pp. 197–202

Refinement of the Otoneurological Decision Support System and Its Knowledge Acquisition Process

Kirsi VARPA[a,1], Kati ILTANEN[a], Martti JUHOLA[a], Erna KENTALA[b] and Ilmari PYYKKÖ[c]
[a] *University of Tampere, Tampere, Finland*
[b] *Helsinki University Central Hospital, Helsinki, Finland*
[c] *Tampere University Hospital, Tampere, Finland*

Abstract. In this paper, we present an otoneurological decision support system ONE and describe its state after upgrade process. Upgrade involved further development of user interface, creation of methods for data transfer and refinement of knowledge base. First we asked physicians to update the knowledge base of the system. For knowledge refinement we developed also a machine learning method that discovers knowledge from data. Refined knowledge bases were tested with otoneurological data. Test results showed that experts' knowledge combined with machine learnt knowledge had the best classification accuracy. The result of the upgrade process is a more usable decision support system.
Keywords: Clinical Decision Support Systems, Otoneurology, Machine Learning.

1. Introduction

Vertigo can be a symptom of many different diseases. In the general population of Finland, 29 % of people reported having an experience of vertigo together with a moving sensation [1]. During one year, 5 % of the general population was diagnosed to suffer from vestibular vertigo [2]. Separation of vertigo diseases from each other can be difficult because of their similar symptoms. Also the diagnostic work-up for vertigo is extensive falling into different specialties and, therefore, requiring a vast amount of knowledge to be successful. A decision support system eases the management of information necessary for the work-up. It ensures that all the crucial questions are asked, and, thus, a diagnosis is not based only on partial information. A correct diagnosis is essential for the choice of a proper treatment, which is in some cases even a destructive surgery.

We have been interested in developing a decision support system for the field of otoneurology since the beginning of the 1990s. An Otoneurological Expert System ONE [3] is an academic software tool developed to support decision making and data gathering for diseases involving vertigo. ONE has been shown to be competitive to experts [4] and another otoneurological expert system [5]. It is currently used at

[1] Kirsi Varpa, Department of Computer Sciences, University of Tampere, FIN-33014 University of Tampere, Finland. Kirsi.Varpa@cs.uta.fi

medical education in Helsinki University Central Hospital. Furthermore, we have collected data of 1478 vertiginous patients with the help of ONE. The vertigo data have been utilised in medical analysis and in various experiments (e.g. [6]) concerning machine learning methods [7] such as decision tree induction, genetic algorithms, nearest neighbour classification and neural networks.

Experiences gained from using ONE in medical education, data collection and machine learning research showed that the system needs and is worthwhile of further development. During the upgrade process we built more user-friendly graphical user interface containing navigation tree, restructured database enabling patient history, refined knowledge base with experts and machine learning method, and made tools for data collection ensuring data quality [8]. In this paper, we present the state of the system after these upgrades.

2. Description of ONE

ONE is implemented in the Java programming language. The main components of the system are the graphical user interface, inference engine, knowledge base, query base and answer database (Figure 1). The query base contains instructions to create the user interface. The instructions tell what questions are shown at the same time in the query panel, what kind of questions a user is dealing with, and what are the possible answers to the questions. There are 192 questions in the query base concerning symptoms, medical history, clinical findings and life quality. In the symptoms part, questions concern, for example, vertigo, gait difficulties, hearing loss and tinnitus. The medical history contains questions about drug usage, head and ear injury, ear operations and other diseases. In the clinical findings, the results of different tests (i.e. otoneurologic, audiologic and imaging tests) are given.

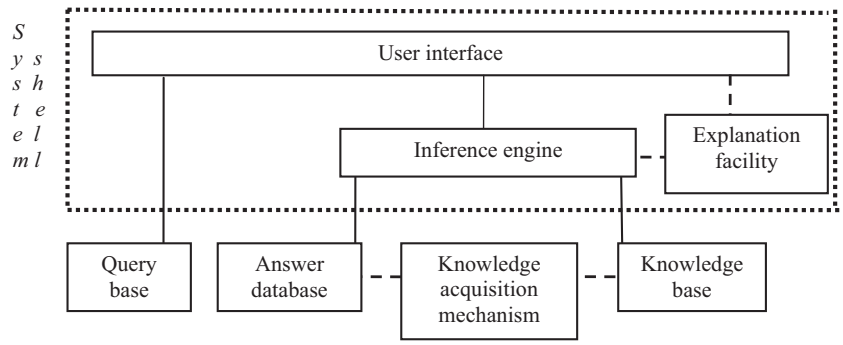


Figure 1. Main components of the otoneurological decision support system One.

The graphical user interface is divided into two different parts: an upper panel and a lower panel. In the upper panel personal information is asked about the patient and the three best diagnosis suggestions by ONE are shown. The lower panel contains a navigation tree and a query panel. The navigation tree shows in which part of the questionnaire the user is. The user can easily see from the leaf symbols of the navigation tree which parts of the questionnaire require more attention: A leaf containing question mark shows that there are unanswered questions in that part of the query. An exclamation mark tells that a patient has answered positively some of the

questions. If the leaf is empty, the patient has not answered positively in any of the questions in the section, in other words, she does not have the symptoms at issue.

The system stores users' answers to a MySQL database. The database contains 35 tables. One table of the database corresponds to one subject of matter in the questionnaire, for example hearing loss or tinnitus. Usually one query view of the system is one table in the database. The database is planned to minimize the number of missing values and to avoid redundancy of data. The database also allows saving the patient's history so that it is possible to follow the evolution of patient's symptoms and to trace the changes in those.

The knowledge base contains a description or a pattern for each disease in the form of fitness values and weights. The current version covers 15 diseases and disorders: acoustic neurinoma, autoimmune disease, benign positional vertigo, benign paroxysmal vertigo of childhood, benign recurrent vertigo, borreliosis, brain stem ischemia, central nervous system (cns) tumour, Menière's disease, ototoxicity, perilymphatic fistula, sudden deafness, traumatic vertigo, vertebrobasilar insufficiency and vestibular neuritis. A weight value assigned to an attribute expresses the significance of the attribute for a disease. Fitness values set to attribute values express the correspondence between these values and the disease. The number of relevant attributes varies according to the diseases: Some diseases can be inferred with few attributes, other diseases require a larger number of attributes. In addition to basic attributes, the disease patterns contain derived attributes computed from the original attributes by logical and arithmetical operations.

The inference method resembles the nearest neighbour method [7], but instead of the nearest neighbour, it searches the nearest patterns (diseases). Furthermore, ONE uses rules concerning necessary attribute values defined for diseases. If a case being classified does not conform to these values, the system tells this for the user through explanation facility. The system tells also if some crucial information is missing. The inference engine counts scores for all the diseases based on the answers given by the user. The diseases with the highest scores are the best fits and suggested by ONE. The engine also counts minimum and maximum scores for diseases taking into account all of the questions, also the ones that the patient has not answered. For the minimum score the inference engine takes into account the lowest fitness values of the unanswered attributes in addition to answered questions and, for the maximum score, the highest fitness values of unanswered questions.

3. Data transfer

Reliable transfer of patient information from the otoneurological questionnaire to the database is important because it affects the quality of the data that, in its turn, has an effect on the results of machine learning methods. Automatic data transfer ensures the quality of data by minimizing the input errors that can occur when a manual input is used. The manual input takes quite a long time, because there are nearly a hundred questions in a questionnaire. Physicians perceive manual data transmission problematic, because there are simply no resources in organisations to do that [9]. In order to ease the data transmission into the system, we developed two different methods to do that.

A way to ease the data transfer is scannable paper questionnaires. The questionnaires are made with Snap Survey Software (SNAP), and paper forms filled by

patients are scanned in it with a text scanner. Scannable questionnaires improve the transmission of information to electronic form. However, scanning cannot be done without personnel. Someone has to input the questionnaires into the scanner and also check that it has read answers correctly. She also has to alter the data collected with SNAP to the right value ranges before transferring it into the database of ONE.

The other solution for the data transmission is a web questionnaire. Web form uses the same database as ONE, and, thus, collected information can be used in real time. Web form is a better solution than scannable questionnaires, because it does not require data alteration: Data is in the right form from the beginning.

4. Knowledge refinement

A part of the upgrade process of the otoneurological decision support system was the refinement of its knowledge base [8]. First we updated original knowledge base with the help of otoneurological experts. They went through the patterns of diseases and updated the weights and fitness values of the attributes, if necessary. After the experts' update process we used a machine learning method [10] to formulate the fitness values for different diseases. The machine learning method is based on the frequency distributions of the attributes. It was possible to create patterns only for the seven most frequent diseases that had enough example cases for the knowledge calculation. We also combined weight values given by the experts to fitness values calculated from the patient data.

After formulating different knowledge base combinations, we tested the knowledge bases with originally collected patient data. Results of the three test drives with original data are shown in Table 1. The results show the classification accuracies of the knowledge bases within the first diagnosis suggestion and the three first diagnosis suggestions.

Table 1. Classification accuracies of different knowledge bases.

		1. Diagnosis Suggestion [%]				1., 2. and 3. Diagnosis Suggestion [%]			
Disease	**Cases**	**Old Know-ledge**	**New Expert Know-ledge**	**ML Know ledge**	**Expert and ML Know-ledge**	**Old Know-ledge**	**New Expert Know-ledge**	**ML Know-ledge**	**Expert and ML Know-ledge**
Acoustic Neurinoma	131	8,4	31,3	66,4	16,8	32,1	68,7	87	71,8
Benign Positional Vertigo	173	5,2	56,7	35,3	50,3	50,9	93,6	64,2	89,0
Menière's Disease	350	22,3	36,0	83,1	77,1	72,9	76,0	96,9	98,3
Sudden Deafness	47	48,9	72,3	74,5	87,2	91,5	93,6	97,9	100,0
Traumatic Vertigo	73	65,8	65,8	78,1	52,1	95,9	98,6	98,6	94,5
Vestibular Neuritis	157	6,4	14,7	65	66,9	84,7	84,7	75,8	93,0
Benign Recurrent Vertigo	20	20,0	70,0	40	50,0	95,0	100,0	85	95,0
Sum	951	19,2	40,4	67,4	60,3	68,4	82,8	86,0	91,8

It can be seen clearly that the update of the knowledge of ONE was necessary. The updated knowledge bases classified cases notably better than the original knowledge used by ONE. The classification accuracy of the original knowledge base as the first diagnosis suggestion was 19.2 %, when the updated knowledge bases classified 40.4–67.4 % of cases correctly. The knowledge base combining expert and machine learning knowledge was generally better than the new knowledge base using only expert knowledge. The classification accuracy with combined knowledge was 60.3 % and with experts' knowledge 40.4 %, when looking at the first diagnosis suggestions. When examining the three diagnosis suggestions, it can be seen that the combined knowledge base is also better than the knowledge base formed only with the machine learning method.

5. Results

The result of the upgrade process of the decision support system is a more user-friendly system. Its graphical user interface, especially the navigation tree, helps the user to easily notice parts of the questionnaire that require extra attention. The user does not have to go through the whole questionnaire any more to form a picture of a patient's symptoms. Instead, she just has to look at the navigation tree.

Data transfer has been altered to a more automatic form, which makes the system also more usable. Patient information from paper questionnaires can be transferred to electronic form more easily with text scanners. This still requires data alteration. When the web questionnaire saves data to the database of the decision support system, data can be used at real time, and, thus, the system can offer more support for physicians when they are diagnosing the case.

The knowledge refinement affected positively to the inference capability of the system. Experts' knowledge update doubled the mean classification accuracy and machine learning (ML) improved it more than treble when looking at the first diagnosis. Experts' attribute selection and weighting is necessary. Generally, combining weights set by the experts to the fitness values calculated by the machine learning method resulted in the best classification results.

6. Discussion

Vertigo diseases can be difficult to separate from each other because of their similar symptoms. Therefore, physicians see systems that help making diagnosis very useful [9]. ONE is planned to support physicians' diagnostic process. It infers a patient's possible diagnoses on the basis of information given about the patient. It also tries to explain why inferred diseases are possible and informs if something important information is missing. Physicians can use it as an assistant when diagnosing patients. With the help of ONE it is possible to collect data systematically for medical research. Collected data can also be utilised in machine learning research which produces methods for finding models describing an application area.

In this study, the combination of experts' knowledge and machine learnt knowledge yielded the best classification accuracies. However, the accuracies of the first diagnosis suggestions are still quite low. A reason for this might be the difficulty of the application area. Some cases are especially difficult to diagnose because of the

phase of the disease. In order to overcome this problem, the knowledge refinement process will continue with the development of a knowledge acquisition tool containing machine learning capabilities. During the upgrade process, we developed a method for formulating fitness values from the data. Later we are going to add weight calculation to the machine learning method. The architecture of ONE with the separate query base, patient database, knowledge base and inference mechanism enables its customising for different institutions in the field of otoneurology and even in different application areas. The knowledge acquisition tool will also help in the adaptation of the system to other institutions.

Acknowledgments

The authors wish to thank Matti J. Tapani, Sari Mykkänen, Minna Kokkonen, Pia Lindberg, Ville Mäkiranta, Eeva Korhonen, and Kalle Mäkelä for their aid to the study. The first author acknowledges the support of the Academy of Finland (grants 78676, 104791 and 202185) and the European Commission (QLRT-2001-02705).

References

[1] Havia M, Kentala E, Pyykkö I. Prevalence of Menière's disease in general population of Southern Finland. Otolaryngol Head & Neck Surg. 2005;133(5):762-8.
[2] Neuhauser HK, von Brevern M, Radtke A, et al. Epidemiology of vestibular vertigo: a neurotologic survey of the general population. Neurology. 2005;65(6):898-904.
[3] Auramo Y. Construction of an expert system to support otoneurological vertigo diagnosis [dissertation]. Tampere: University of Tampere; 1999.
[4] Kentala E, Auramo Y, Juhola M, Pyykkö I. Comparison between diagnoses of human experts and a neurotologic expert system. Ann Otol Rhinol Laryngol. 1998;107:135-40.
[5] Auramo Y, Juhola M. Comparison of inference results of two otoneurological expert systems. International Journal of Bio-Medical Computing. 1995;39:327-35.
[6] Juhola M, Laurikkala J, Viikki K, Auramo Y, Kentala E, Pyykkö I. Neural network recognition of otoneurological vertigo diseases with comparison of some other classification methods. In: Horn W, Sharar Y, Lindberg G, Andreassen S, Wyatt J, editors. Artificial Intelligence in Medicine, volume 1620 of Lecture Notes in Computer Science. Berlin: Springer; 1997. p. 217-26.
[7] Mitchell T. Machine Learning. New York: McGraw-Hill; 1997.
[8] Varpa K. Knowledge representation and acquisition in knowledge-based systems and the renewal of the otoneurological decision support system and its knowledge (in Finnish) [master thesis]. Tampere: University of Tampere; 2005.
[9] Aalto P. Market account of Equihear. IT-based concept of hearing and vertigo (in Finnish). Tampere: Finn-Medi Tutkimus; 2005.
[10] Viikki K, Juhola M. Refining the knowledge base of an otoneurological expert system. In: Crespo J, Maojo V, Martin F, editors. Medical Data Analysis, volume 2199 of Lecture Notes in Computer Science. Berlin: Springer; 2001. p. 276-81.

PUBLICATION II

Machine Learning Method for Knowledge Discovery Experimented with Otoneurological Data

Kirsi Varpa, Kati Iltanen and Martti Juhola

Computer Methods and Programs in Biomedicine 91(2), 2008, pp. 154–164
https://doi.org/10.1016/j.cmpb.2008.03.003

COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 91 (2008) 154–164

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Machine learning method for knowledge discovery experimented with otoneurological data

Kirsi Varpa, Kati Iltanen, Martti Juhola*

Department of Computer Sciences, FI-33014 University of Tampere, Tampere, Finland

ARTICLE INFO

Article history:
Received 23 January 2007
Received in revised form
17 March 2008
Accepted 18 March 2008

Keywords:
Knowledge discovery
Machine learning method
Otoneurology

ABSTRACT

We have been interested in developing an otoneurological decision support system that supports diagnostics of vertigo diseases. In this study, we concentrate on testing its inference mechanism and knowledge discovery method. Knowledge is presented as patterns of classes. Each pattern includes attributes with weight and fitness values concerning the class. With the knowledge discovery method it is possible to form fitness values from data. Knowledge formation is based on frequency distributions of attributes. Knowledge formed by the knowledge discovery method is tested with two vertigo data sets and compared to experts' knowledge. The experts' and machine learnt knowledge are also combined in various ways in order to examine effects of weights on classification accuracy. The classification accuracy of knowledge discovery method is compared to 1- and 5-nearest neighbour method and Naive–Bayes classifier. The results showed that knowledge bases combining machine learnt knowledge with the experts' knowledge yielded the best classification accuracies. Further, attribute weighting had an important effect on the classification capability of the system. When considering different diseases in the used data sets, the performance of the knowledge discovery method and the inference method is comparable to other methods employed in this study.

1. Introduction

In the general population of Southern Finland, 29% of people reported having an experience of vertigo together with a moving sensation [1]. In Germany, 5% of the general population was diagnosed to have vestibular vertigo within a year [2]. Because vertigo is the principal symptom in most of the otoneurological disorders having also other similar symptoms [3,4], it is difficult to separate disorders from each other. A correct diagnosis and treatment of a disorder can prevent a patient to be isolated and handicapped, at best avoid a destructive surgery or even a death caused by falling [3]. We have been developing an otoneurological decision support system ONE [5–9] during the last decade in order to support diagnostics of vertigo diseases. Its purpose is to aid a decision maker by offering diagnosis suggestions inferred from the information gathered on a patient.

The architecture of the decision support system ONE is separate: it consists of different components connected together [5]. The components are the query base, graphical user interface, knowledge base, answer database, inference mechanism, and explanation facility. The separate architecture of ONE enables its customizing for different institutes relatively easily, because only the query base, knowledge base and answer base have to be tailored. Its graphical user interface, inference method and explanation facility form a kind of expert system shell that can be used as a basis in creating new decision support systems even for different application domains.

* *Corresponding author.*
E-mail address: Kirsi.Varpa@cs.uta.fi (K. Varpa).
0169-2607/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2008.03.003

The knowledge base of ONE was originally formed by domain experts [6]. There occurred some problems in the separation of disease classes when using this knowledge in inference and, therefore, the refinement of the knowledge base was started [7] and a machine learning (ML) method, that creates knowledge by analysing domain data, was developed [7]. In [7], an otoneurological data set of 815 cases collected in the Department of Otorhinolaryngology at Helsinki University Central Hospital in Finland was used. Gathering this data set was started for the development of ONE in the beginning of 1990s and it continued during the decade [8]. The cases of the data set belong to the diagnostic groups of acoustic neurinoma, Ménière's disease, benign positional vertigo (BPV), sudden deafness, traumatic vertigo, and vestibular neuritis. Results of [7] suggested that knowledge learnt from data was useful when refining the knowledge base. However, there were some limitations in the study. One limitation was the approach used to estimate the predictive performance of ONE with the knowledge learnt from the data: the knowledge was learnt from 70% of cases and it was tested with the rest 30% of the cases. Other limitation was the lack of comparison methods, i.e., the method was not compared to other ML methods.

The decision support system ONE was upgraded in order to achieve a more user-friendly graphical user interface, to restructure its database and to make data transferring tools ensuring data quality [9]. During the renewal process, ONE's query and knowledge bases were upgraded. Some new important questions were added into the query base and some of the old questions and their answer alternatives were modified to make them more understandable. The upgrade of the knowledge base was first done manually with otoneurological experts who updated the disease patterns when necessary. The experts used their knowledge and experience as the basis when updating the knowledge base. They had also a possibility to compare their assumptions about the diseases with the collected data. After that, another knowledge base was formed with the ML method [7]. Further, a combination of experts' and machine learnt knowledge was created for preliminary testing. Results showed that knowledge upgrade was necessary and enhanced notably the classification accuracy of the decision support system. Still, there seemed to be some difficulties in recognition of certain disease classes (acoustic neurinoma and Ménière's disease).

In this paper, we describe the ML method developed in [7] and test its functionality with ONE's inference. We compare classification accuracies of the knowledge created by the ML method to the experts' knowledge and to various combinations of expert and machine learnt knowledge. To estimate the predictive performance of machine learnt knowledge, we use the 10-fold cross-validation [10], which is a more sophisticated evaluation method than the pure training set/test set division and gives more reliable results with data sets of a small or moderate size [11]. In the test, we use an augmented vertigo data set of 1030 cases, which has additional cases collected at Helsinki University Central Hospital in the beginning of the 2000s. In addition to cases from the previously mentioned six classes, the extension has cases from three new classes: benign recurrent vertigo (BRV), vestibulopatia and central lesion. Further, a totally new data set of 253 cases collected at Tampere University Hospital in Finland is used to evaluate the knowledge learnt from data. Finally, we compare results of ONE's inference to results of the *k*-nearest neighbour (*k*-NN) method [10] and Naive–Bayes (NB) classifier [10,12] to test its general functionality.

2. Machine learning

ML methods [10] can be used in knowledge refinement and knowledge discovery if there is enough data collected from a domain. For example, with ML methods it is possible to try to discover hidden patterns or rules occurring in data. Some of the methods can give explanations about the way of achieving results, thus, enabling the user to evaluate the formed results [5]. Generally used knowledge discovery methods are decision tree induction [13], *k*-nearest neighbour method [10], NB classifier [12,10], and neural networks (NNs) [10].

Decision tree algorithms use inductive inference: they try to build general models from the example cases using heuristics, e.g., ID3 [13], C4.5 [11]. Each path from the root node of the decision tree to a leaf node can be regarded as a rule that leads to a certain class through different attribute tests (nodes). Decision trees are quite easy to understand and, maybe because of this, they have been used widely in different domains. Decision trees have been applied in medicine, e.g., in oncology to manage decision protocols [14], to create classification patterns of metabolic disorders [15], to acquire knowledge about otoneurological diseases [8], in data mining of diabetes [16,17], and to distinguish the severity of dementia [18].

The *k*-nearest neighbour method is an example of instance-based learning [10]. It tries to find the most similar cases of the new case from the learning data (example cases) by using some distance metric. The number of *k* tells how many neighbours (similar cases) are searched. The class of the new case is the class being in the majority among the *k*-nearest cases. The *k*-nearest neighbour method has been used, e.g., in generating a system for evidence-based medicine in oncology [19], in medical information retrieval [20], in bioinformatics to predict the protein β-turn [21] and in gene selection from microarray data [22], and in texture analysis in breast tissue analysis and characterization [23].

The NB classifier is a classification method that is based on probability calculation [10,12]. It assumes that all the attribute values are conditionally independent, which simplifies its calculation. A new case is classified into the class having the highest calculated probability. The NB classifier has been utilised, e.g., in feature selection and classification model construction of diabetes data [16], for classifying oncology [24], for predicting microRNA genes [25] and target [26], and predicting survival rate of cirrhotic patients [27].

NNs consist of processing units, neurons that are connected together in input, hidden and output layers [10]. They try to determine the proper weights on the interconnections between neurons in order to achieve correct output. NNs have been used in medicine, for example, in diagnosing heart murmurs in paediatrics [28] and unbalance with acoustic signals [29], classifying otoneurological patients by eye movement signals [30] and in bioinformatics to predict solvent accessibility

of amino acid residues in proteins [31], and sequence of TP53 gene [32].

3. Description of inference mechanism and knowledge discovery method

The decision support system ONE infers diagnosis suggestions about the given information of the subject [5]. Its inference mechanism resembles pattern recognition methods, especially the weighted k-nearest neighbour method [33]. Inference is based on the weight and fitness values set for attributes, which are referred to as the knowledge of the system.

3.1. *Knowledge base*

The knowledge base of ONE contains a description or a pattern for each deducible class (disease) in the form of fitness values and weights [5]. A weight value assigned to an attribute expresses the significance of the attribute for the class. Weight values vary from 0 to chosen maximum, where 0 means that the attribute does not concern the class at all. The greater

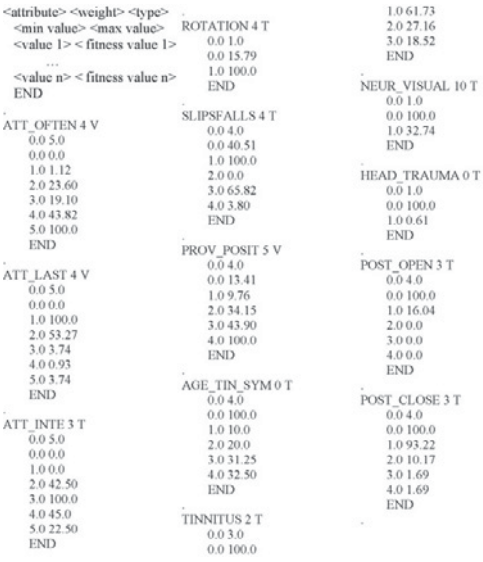


Fig. 1 – Part of knowledge pattern for benign positional vertigo. Meaning of attributes: ATT_OFTEN = frequency of vertigo attacks, ATT_LAST = length of the vertigo attacks, ATT_INTE = severity of vertigo attacks, ROTATION = feeling of rotation, SLIPSFALLS = frequency of Tumarkin-type drop attacks, PROV_POSIT = severity of position induced vertigo, AGE_TIN_SYM = occurrence of tinnitus, TINNITUS = handicap caused by tinnitus, NEUR_VISUAL = visual blurring or double vision during vertigo attacks, HEAD_TRAUMA = direct injury of the head or neck associated with the beginning of vertigo symptoms, POST_OPEN = posturography base line, eyes open (cm/s), POST_CLOSE = posturography base line, eyes closed (cm/s).

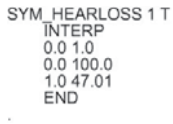


Fig. 2 – Example of an unusual characteristic taken into account in disease pattern of benign positional vertigo. According to the clinical picture of the benign positional vertigo, hearing loss is not typical for this disease. However, many patients do have, e.g., age-related hearing loss.

the weight value is, the more important the attribute is for the class. Fitness values set to attribute values express the correspondence between these values and the class. In the knowledge base, fitness values are set from 0 to 100. The fitness value 0 means that the attribute value does not fit the class whereas the fitness value 100 shows that the value at issue fits best the class. Fig. 1 presents an example of a disease pattern.

In the knowledge base each pattern corresponds to one vertigo disease (class). Disease patterns can be considered as profiles of diseases. These patterns match partly the clinical pictures of diseases. However, in the patterns it is possible to take into account also unusual characteristics that a patient may have. For instance, a patient suffering BPV can have hearing loss because of his age although hearing loss is not normally a symptom of this disease. An example of this is shown in Fig. 2. Due to fitness values, it is possible to take into account this kind of uncharacteristic symptoms.

3.2. *Inference*

The inference method of ONE resembles weighted k-nearest neighbour methods of pattern recognition [33]. Instead of searching the nearest neighbours it searches the fittest disease classes. It calculates scores for the classes from the weight and fitness values of attributes. The score $S(d)$ for the disease d is calculated in the following way

$$S(d) = \frac{\sum_{a=1}^{A(d)} x(a) w(d,a) f(d,a,j)}{\sum_{a=1}^{A(d)} x(a) w(d,a)}, \quad (1)$$

where $A(d)$ is the number of the attributes associated to disease d, $x(a)$ is 1, if the value of attribute a for the disease d is known, otherwise 0, $w(d,a)$ is the weight of the attribute a for the disease d, and $f(d,a,j)$ is the fitness value for the value j of the attribute a for the disease d. In the case of quantitative attributes, the fitness values are interpolated by using attribute values in the knowledge base as interpolation points. The fitness values are altered to the range of 0–1 during the inference process.

The disease class having the highest score is the best diagnosis suggestion. ONE also counts the minimum and maximum scores for the classes using the lowest and the highest fitness values for the attributes having missing values. With the minimum and maximum scores ONE tries to handle uncertainty caused by missing values. The closer the minimum and

maximum scores are to each other, the more reliable the inference is.

The diagnosis suggestions of ONE are ordered primarily by the score and secondarily by the difference of the minimum and maximum score. If the classes have the same score but one class has a smaller difference between the minimum and maximum scores than the others, the class having the smallest difference is placed as a higher diagnosis suggestion. If the classes have the same score and the minimum and maximum score difference, their order is selected randomly.

In addition to diagnosis suggestions, ONE can explain why a class may not be possible on the basis of its knowledge. In the knowledge base, some attributes are marked as necessary attributes. (The type "V" in a knowledge pattern means that an attribute is a necessary one. For ordinary attributes, the type is "T".) If a patient does not have a certain value for a necessary attribute, the system informs the user about this. Due to necessary attributes, ONE can inform if some crucial information about the patient is missing.

3.3. *Learning fitness values from data*

Weights and fitness values in the original knowledge base were set by domain experts on the basis of their experience and knowledge [6]. The weight values varied typically from 0 to 5 but there were also some larger values. The number of relevant attributes in the patterns varied according to the classes: some classes can be inferred with few attributes, other classes require a larger number of attributes.

If the number of diseases and attributes is large, the manual definition of the knowledge base is difficult and tedious. With the knowledge base defined by experts, acoustic neurinoma, BPV, Ménière's disease and vestibular neuritis were not recognized as well as expected [34]. We developed the knowledge discovery method [7] that calculates fitness values for values of attributes from domain data. Fitness values can be perceived as values that show how often values of an attribute occur in a certain class. Fitness values are calculated for every class separately from example cases belonging to it. On the basis of these cases, a frequency distribution is formed for each attribute defined relevant for the class.

The frequency distributions form the basis of the knowledge discovery method. The most frequently occurring attribute value fits best the class. Thus, the fitness value for the attribute value with the highest frequency is set to 100. Fitness values for the other attribute values are formed by relating their frequencies to the frequency of the most frequently occurring value:

$$\mathrm{fv}(d, a, j) = \frac{\mathrm{fr}(d, a, j)}{\mathrm{fr}(d, a, h)} \times 100, \quad (2)$$

where fv(d,a,j) is the fitness value of the value j of the attribute a in class d, fr(d,a,j) is the frequency of the value j of the attribute a in class d and fr(d,a,h) is the highest frequency in the distribution of the attribute a in class d.

The method gets as an input a list of all possible values of qualitative attributes occurring in the domain. If there are attribute values that do not occur in the frequency distribution under consideration, their fitness values are set to 0. Quantitative attributes have to be discretised in order to calculate the fitness values. In the inference, values of new cases can be discretised or fitness values for new cases can be interpolated by using mid-points of discretised intervals as interpolation points.

Table 1 – Frequency distribution of severity of vertigo attacks in benign positional vertigo including machine learnt fitness values

Value	Frequency	%	Fitness value
0 = no vertigo attacks	0	0	0.0
1 = mild	0	0	0.0
2 = weak	34	19.7	42.5
3 = moderate	80	46.2	100.0
4 = strong	36	20.8	45.0
5 = very strong	18	10.4	22.5
Valid	168	97.1	
Missing	5	2.9	
Total	173	100.0	

Let us use cases of BPV and an attribute *severity of the vertigo attacks* (*att_inte*) as an example of the use of the knowledge discovery method. There are 173 patients diagnosed to have this disease. The frequency distribution of the severity of the vertigo attacks in the class BPV is shown in Table 1. It can be seen that value 3 (*moderate*) has the highest frequency of 80, so, it fits best the disease. Therefore, its fitness value is set to 100 in the knowledge base. Other fitness values are calculated by the Eq. (2). For example, the frequency of the attribute value "4 = *strong*" is 36, and, thus, its fitness value is (36/80) × 100 = 45.0. The attribute severity of vertigo attacks has actually six possible attribute values but the first two values "0 = *no vertigo attacks*" and "1 = *mild*" do not occur in the frequency distribution under consideration, and, therefore, their fitness values are set to 0.

4. Results of the knowledge discovery method and inference method of ONE

For this study, we formed knowledge bases with the help of the knowledge discovery method besides the experts. At the moment, the method can calculate only the fitness values for attribute values and sets all the weight values to one. Further, we combined the knowledge of the experts with the knowledge formed by the knowledge discovery method into a knowledge base where weight values are set by the experts and fitness values are calculated from the domain data. The previous study [9] showed that there seemed to be some difficulties in recognition of certain disease classes (acoustic neurinoma and Ménière's disease) when using experts' attribute weighting. Therefore, we tested also how changing the weight values to one affected classification of these diseases. Within this study five different knowledge bases were used:

KB1 = Fitness and weight values were set by the experts. Weight values were set from 0 to 10, where 0 means that an attribute does not concern the class at all and, therefore, it was not taken into account in inference.

KB2 = All weight values were set to 1 and fitness values were calculated from data by the knowledge discovery method.
KB3 = Weight values were set by the experts and fitness values were calculated from data by the knowledge discovery method.
KB4 = As KB3 but weight values of acoustic neurinoma were set to 1 in order to see influence of the weight values in classification on acoustic cases.
KB5 = As KB4 but weight values of Ménière's disease were also set to 1.

The knowledge bases were tested with the augmented vertigo data and with the new vertigo data. To characterize the performance of the knowledge bases, true positive rates (TPRs) and total classification accuracies were calculated. For each disease class, a TPR is calculated as the percentage of correctly inferred cases in the class:

$$\text{TPR } (\%) = 100\frac{tpos}{ncases}, \tag{3}$$

where *tpos* is the number of correctly classified cases in the class and *ncases* is the number of all cases in the disease class. For each knowledge base, a total classification accuracy (ACC) is calculated:

$$\text{ACC } (\%) = 100\frac{t}{n}, \tag{4}$$

where t is the number of cases correctly classified in all disease classes and n is the total number of vertigo cases used in classification.

4.1. *Results with the augmented data*

For testing the knowledge discovery method, the augmented otoneurological data containing 1030 vertigo cases is used. The augmented data contains cases of nine vertigo diseases:

1. acoustic neurinoma ($n_{ane} = 131$; 12.7%),
2. benign positional vertigo ($n_{bpv} = 173$; 16.8%),
3. Ménière's disease ($n_{men} = 350$; 34.0%),
4. sudden deafness ($n_{sud} = 47$; 4.6%),
5. traumatic vertigo ($n_{tra} = 73$; 7.1%),
6. vestibular neuritis ($n_{vne} = 157$; 15.2%),
7. benign recurrent vertigo ($n_{brv} = 20$; 1.9%),
8. vestibulopatia ($n_{ves} = 55$; 5.3%) and
9. central lesion ($n_{cl} = 24$; 2.3%).

Each case is a patient who has informed to have vertigo or gait difficulties and is diagnosed to have one of the listed vertigo diseases. Patients have filled out an otoneurological questionnaire, which has 105 questions concerning symptoms and medical history. Furthermore, there are 72 questions (attributes)[1] about otoneurologic, audiologic and imaging tests in the query base. The tests are not done to every patient, and, thus, for most of the patients there are no answers concerning test results. Attributes with low frequencies of available values are not used in testing knowledge discovery method. In tests, 89 attributes are used: 83 basic attributes and 6 attributes derived from the basic attributes. Quantitative attributes, e.g., age when the symptoms started, results of caloric tests, posturography and audiometry frequency and speech, are discretised into equal-width intervals in order to guarantee a similar way of their handling in all the ML methods used in this study. Thus, all of the attributes used in this study are qualitative. For most of the attributes, the percentage of missing values is about 10. Ten attributes have about 30% of their values missing, and for one important attribute even 53% of values are missing.

The renewal process of the decision support system included update of the otoneurological query base. Some new important questions, e.g., vertigo type, hyperacusis and its handicap, were added into the query base. Some of the old questions and their answer alternatives were modified in order to simplify them and make them easier to answer. Modification concerned mainly the answer alternatives: answer scales were limited to five alternatives. In some cases this meant limitation of alternatives but in other cases new alternatives were added. In order to test knowledge bases with the augmented and the new data, the augmented data had to be transformed into a form matching the new query base. Also, some of the answer alternatives of the new data had to be combined into one option because there were no similar alternatives in the augmented data.

In order to get estimates for the predictive performance of the knowledge bases KB2–KB5 created from the augmented data, a 10-fold cross-validation [10] was used. The augmented data containing 1030 cases was divided into 10 subsets. A knowledge base was created on the basis of nine subsets and its validity was tested with the remaining subset. For example, cases of Ménière's disease (350) were divided into 10 subsets each containing 35 cases. Thus, the pattern of Ménière's disease in the knowledge base was created on the basis of 315 cases and tested with 35 cases. This was repeated 10 times altogether. The classification suggestions for each test set were output into one file, and TPRs and total classification accuracies were calculated directly from correctly classified cases. The 10-fold cross-validation was done three times using three different random data divisions. For all the knowledge bases, the same three data divisions were used for the sake of comparability.

There were approximately 10 cases with the same scores and score differences in the results yielded by most of the knowledge bases. Most of the cases with the same scores and differences occurred in the results of the KB2 where weight values were equal to 1: the first and second diagnosis suggestions had the same score and score difference in 90 cases, and 12 cases had the same score and score difference even in first, second and third diagnosis suggestions. The order of the three first diagnosis suggestions having the same score and score difference was selected randomly likewise in the case of two diagnosis suggestions having the same score and difference.

In the result Tables 2–4, mean TPRs and classification accuracies of the three 10-fold cross-validations for knowledge bases KB2, KB3, KB4 and KB5 are presented. Error bars (99% confidence intervals) for the results of each cross-validation

[1] A question can be thought as an attribute, and, thus, we use the term attribute instead of question henceforth.

Table 2 – True positive rates of disease classes and total classification accuracies of knowledge bases within the first diagnosis suggestion in percents in augmented data

Disease name	Cases	KB1	KB2	KB3	KB4	KB5
Acoustic neurinoma	131	31.3	63.4	15.5	69.7	62.1
Benign positional vertigo	173	59.5	27.4	44.1	44.5	40.3
Ménière's disease	350	38.3	80.7	76.7	74.2	91.5
Sudden deafness	47	72.3	66.0	84.4	84.4	75.2
Traumatic vertigo	73	68.5	66.2	41.6	41.6	42.0
Vestibular neuritis	157	18.5	61.8	66.2	66.9	65.2
Benign recurrent vertigo	20	60.0	10.0	30.0	30.0	21.7
Median of TPR		59.5	63.4	44.1	66.9	62.1
Total accuracy	951	42.4	62.1	57.3	64.0	67.7

are shown in Figs. 3 and 4. When looking at the total classification accuracies of the first diagnosis suggestions (Table 2), it can be seen that the knowledge bases using machine learnt knowledge classify cases more accurately than the knowledge base defined by experts (42.4%). The best ACC (67.7%) is gained by KB5 that combines the knowledge formed by the discovery method with experts' knowledge, in which weight values of acoustic neurinoma and Ménière's disease are set to 1. Overall, accuracies do not differ significantly on the basis of the error bars of Fig. 3. However, the accuracy of KB5 is significantly higher than the accuracy of KB3 in the cross-validations 1 and 2. On the basis of the total

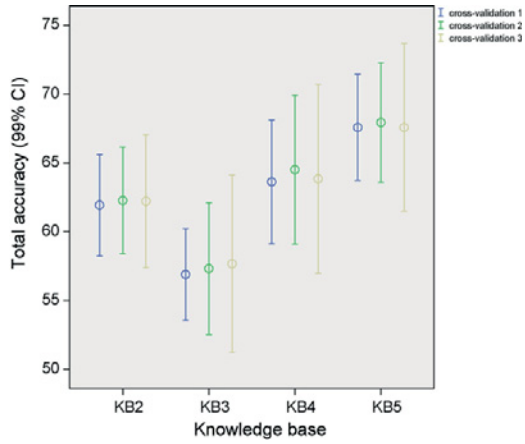


Fig. 3 – Error bars (99% confidence intervals) for total accuracies of each cross-validation with different knowledge bases within the first diagnosis suggestions in augmented data.

Table 3 – True positive rates of disease classes and total classification accuracies of knowledge bases within the first and second diagnosis suggestions in percents in augmented data

Disease name	Cases	KB1	KB2	KB3	KB4	KB5
Acoustic neurinoma	131	45.8	74.6	64.4	90.3	84.2
Benign positional vertigo	173	87.3	47.0	71.1	69.2	65.9
Ménière's disease	350	60.3	94.5	94.0	91.6	97.6
Sudden deafness	47	85.1	86.5	97.9	96.5	94.3
Traumatic vertigo	73	91.8	84.0	79.9	79.9	77.2
Vestibular neuritis	157	52.9	68.4	84.3	83.0	80.9
Benign recurrent vertigo	20	95.0	30.0	60.0	60.0	60.0
Median of TPR		85.1	74.6	79.9	83.0	80.9
Total accuracy	951	66.4	76.2	82.5	84.6	84.7

Table 4 – True positive rates of disease classes and total classification accuracies of knowledge bases within the first, second and third diagnosis suggestions in percents in augmented data

Disease name	Cases	KB1	KB2	KB3	KB4	KB5
Acoustic neurinoma	131	68.7	85.5	71.0	95.2	94.7
Benign positional vertigo	173	94.8	62.4	88.1	84.8	81.7
Ménière's disease	350	76.3	97.0	97.9	97.2	99.8
Sudden deafness	47	93.6	92.9	99.3	97.9	97.9
Traumatic vertigo	73	98.6	95.0	94.5	94.5	93.6
Vestibular neuritis	157	87.9	75.4	92.6	88.7	88.5
Benign recurrent vertigo	20	100.0	46.7	75.0	75.0	75.0
Median of TPR		93.6	85.5	92.6	94.5	93.6
Total accuracy	951	83.6	84.1	90.9	92.6	92.8

classification accuracies of the first and second diagnosis suggestions (Table 3) and of the first, second and third suggestions (Table 4), the knowledge bases combining the machine learnt knowledge and the experts' knowledge classify cases more accurately than the bases completely formed by the experts or the ML method. Within the first and second suggestions, 82.5–84.7% of cases are classified correctly with the knowledge combinations and within the first, second and third suggestions, 90.9–92.8% of cases. With the experts' knowledge, the former classification accuracy is 66.4% and the latter 83.6%.

In spite of the lowest ACC, the knowledge base KB1 using only the experts' knowledge has remarkably better TPRs in

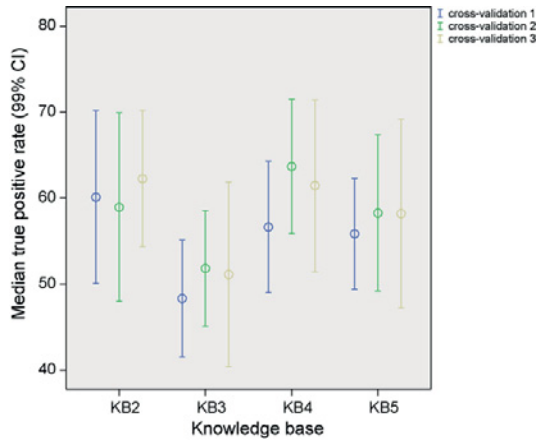


Fig. 4 – Error bars (99% confidence intervals) for median true positive rates of each cross-validation with different knowledge bases within the first diagnosis suggestions in augmented data.

two disease classes, BPV and BRV. For example, the TPR of BPV within the first diagnosis suggestion is 59.5% and within the first, second and third suggestions 94.8%, when in the other knowledge bases the corresponding rates are 27.4–44.5% and 62.4–88.1%, respectively. The knowledge bases using machine learnt knowledge and experts' knowledge do still have some difficulties in separating especially these two classes from others. Traumatic vertigo is also identified better with the pure experts' knowledge. Instead, with other disease classes, especially acoustic neurinoma and Ménière's disease, there seem to be problems in classifying with KB1. When looking at the first diagnosis suggestion, KB1 classifies cases of acoustic neurinoma (31.3%), Ménière's disease (38.3%) and vestibular neuritis (18.5%) remarkably worse than KB2 containing weight values 1 and fitness values calculated from data classify cases (63.4%, 80.7%, 61.8%, respectively). This indicates that there are some problems in experts' setting of weight or fitness values.

The appropriate fitness values are crucial in classification and, as can be seen in Tables 2–4, they alone can improve the classification accuracies. The combination of proper weight and fitness values can improve the classification even more. Therefore, we combined the weight values from the knowledge base KB1 with the fitness values from the knowledge base KB2 and got as a result the knowledge base KB3. The combination did not work as well as expected. The TPR of acoustic neurinoma decreased from 63.4% (KB2) to 15.5% when considering the first diagnosis suggestion and from 74.6% (KB2) to 64.4% within the first and second suggestions. For sudden deafness and vestibular neuritis, the combination improved the TPR but, instead, identification of traumatic vertigo decreased.

In order to find out whether there were really problems in weight values set by the experts, the weight values of acoustic neurinoma were changed to 1 in KB4. The weight alteration improved the TPR of acoustic neurinoma. As the first suggestion, 69.7% of cases are correctly recognized and even 90.3% of cases when looking at the first and second diagnosis suggestion. Because Ménière's disease was not recognized with the experts' weights as well as in knowledge base KB2, we also made the weight alterations to the weights of Ménière's disease (KB5). With KB5, 91.5% of Ménière cases were identified correctly as the first diagnosis suggestion and 99.8% within the first, second and third diagnosis suggestion. The ACC was also the highest in KB5. On the basis of the median TPRs of the first diagnosis suggestion, the performance of KB3 is worst (44.1%) in comparison with the other knowledge bases having machine learnt fitness values. KB4 has the best median TPR (66.9%). However, the error bars of Fig. 4 do not reveal any significant differences.

4.2. *Results with the new data*

The knowledge bases formed from the augmented data were tested also with a totally new test data that contains 310 cases. The test data was gathered during 2004–2005 at Tampere University Hospital. In some disease classes the number of cases was so small that we could not test their classification. Hence, the classes containing over 20 cases were selected to test drives: BPV ($n_{bpv} = 80$), Ménière's disease ($n_{men} = 128$), vestibular neuritis ($n_{vne} = 20$) and vestibulopatia ($n_{ves} = 20$). We could not use vestibulopatia in all the test drives because there was not a knowledge pattern for it in the knowledge base KB1.

The TPRs and total classification accuracies within the new data are shown in Table 5. The best ACC of the first diagnosis suggestion is in KB5. It classifies 53.5% of cases correctly when the other bases get 23.2–46.1% of cases correctly. When looking at the two and three first diagnosis suggestions, KB3 has the highest ACC. It classifies 68.9% and 86.0% of cases correctly when the total classification accuracies for the other knowledge bases are 51.8–68.0% and 59.2–80.3%, respectively. It seems to be that the knowledge bases combining the experts' knowledge and machine learnt knowledge are also more appropriate for the new data. Within the three diagnosis suggestions these knowledge combining bases classify over 79.8% of cases correctly when with the pure experts' knowledge the accuracy is 71.9% and with the pure machine learnt knowledge 59.2%.

From the TPRs of the disease classes in Table 5, it can be seen that the pure experts' knowledge in KB1 does not work with the new data as well as with the augmented data. Especially, BPV cases are difficult to recognize. Only 13.8% of BPV cases are classified correctly as the first diagnosis suggestion and 67.5% as the first, second or third suggestion.

Table 5 – True positive rates of disease classes and total classification accuracies of knowledge bases in percents in new data

	Disease name	Cases	KB1	KB2	KB3	KB4	KB5
First diagnosis suggestion	Benign positional vertigo	80	13.8	1.3	12.5	10.0	12.5
	Ménière's disease	128	29.7	78.9	68.8	65.6	83.6
	Vestibular neuritis	20	20.0	15.0	30.0	25.0	25.0
	Total 1. accuracy	228	23.2	46.1	45.6	42.5	53.5
First or second diagnosis suggestion	Benign positional vertigo	80	41.3	5.0	42.5	42.5	35.0
	Ménière's disease	128	60.2	85.9	88.3	82.0	93.0
	Vestibular neuritis	20	50.0	20.0	50.0	45.0	40.0
	Total 1.–2. accuracy	228	52.6	51.8	68.9	64.9	68.0
First, second or third diagnosis suggestion	Benign positional vertigo	80	67.5	15.0	73.8	62.5	57.5
	Ménière's disease	128	73.4	89.8	96.1	94.5	97.7
	Vestibular neuritis	20	80.0	40.0	70.0	60.0	55.0
	Total 1.–3. accuracy	228	71.9	59.2	86.0	80.3	79.8

Anyhow, this is still much better than with the knowledge formed by ML method in KB2 (1.3% and 15.0%, respectively). With the experts' knowledge, vestibular neuritis cases are also classified more accurately than with KB2. Instead, Ménière's disease is classified notably better by the knowledge formed by the knowledge discovery method. With the knowledge base KB2, 78.9% of cases are recognized as the first suggestion and 89.8% as the first, second and third suggestions, when the corresponding values are 29.7% and 73.4% using the pure experts' knowledge. In general, the combination of the experts' and machine learnt knowledge (KB3) is better than the pure experts' knowledge or the knowledge from the discovery method. Setting weights of acoustic neurinoma to one in KB4 does not enhance classification. Instead, setting also the weights of Ménière's disease to one does improve the recognition of Ménière's cases.

The difference between the total classification accuracies of the augmented and the new data with different knowledge bases ranges between 8.2% and 24.6% when classifying disease classes BPV, Ménière's disease and vestibular neuritis. The knowledge fits better to the augmented data. The mean TPRs of BPV and vestibular neuritis with the different knowledge bases are almost 30% higher with the augmented data compared to the new data. There seem to be some problems with both data in the recognition of BPV cases as the first diagnosis suggestion.

5. Comparison of ONE's inference with the *k*-nearest neighbour method and the NB classifier

To test the classification capability of ONE, its first diagnosis suggestions gained with the knowledge base KB2 were compared to the classifications given by the 1- and 5-nearest neighbour methods and the NB classifier. The *k*-nearest neighbour method and the NB classifier somewhat resemble ONE's inference mechanism. All these methods are simple to compute and easy to interpret. They treat attributes as independent of each other; we used the *k*-nearest neighbour method with the unweighted version of the Value Difference Metric [35]. The theoretical error rate of NB classifier is minimal compared to all other classifiers [10]. In practical domains, the assumption of class conditional independency, for example, is usually violated, and, thus, the performance of NB has been found to be comparable to decision trees and NNs in various empirical studies [10]. This has been shown, for instance, in the study of Kononenko et al. [36]. The error rate of the *k*-nearest neighbour method approaches the Bayes error, when the size of the training data and *k* both approach infinity [10].

The knowledge base KB2 was used in the comparisons because the weights were not used either in distance calculation of the *k*-nearest neighbour methods or in NB classification. In NB classification, the Laplace-estimate [37] was used for estimation of prior probabilities and the M-estimate [10,37] for estimation of conditional probabilities. In the 5-nearest neighbour (5-NN) method, the predicted class was selected randomly if there were more than one majority class within the 5-nearest cases. This is similar to the random way in which ONE's inference selects the order of the classes if they have the same score and score difference. In different cross-validation drives with 5-NN, there were 58–60 cases which had two majority classes and 9–12 cases which had five majority classes.

Table 6 – True positive rates of disease classes and total classification accuracies of ONE for KB2, *k*-nearest neighbour methods and Naive–Bayes classifier in percents in augmented data

Disease name	Cases	ONE (KB2)	1-NN	5-NN	NB
Acoustic neurinoma	131	63.4	75.8	75.6	78.9
Benign positional vertigo	173	27.4	64.5	64.9	58.6
Ménière's disease	350	80.7	88.1	95.9	89.0
Sudden deafness	47	66.0	61.7	51.8	63.8
Traumatic vertigo	73	66.2	80.4	79.0	77.6
Vestibular neuritis	157	61.8	80.0	80.5	78.1
Benign recurrent vertigo	20	10.0	15.0	6.7	11.7
Vestibulopatia	55	37.0	23.6	26.1	9.1
Central lesion	24	37.5	5.6	5.6	22.2
Median of TPR		61.8	64.5	64.9	63.8
Total accuracy	1030	60.2	72.8	75.0	71.7

The methods were tested with the augmented and new data. In the test drives with the augmented data, nine disease classes were used whereas with the new data it was possible to use only four disease classes because there were not enough cases in other classes. In the tests with the augmented data, 10-fold cross validation was employed three times. Mean results of the test drives with the augmented data are shown in Table 6. Error bars (99% confidence intervals) for the results of each cross validation with the augmented data and different ML methods are shown in Figs. 5 and 6. In the test with the new data, the augmented data was used as a learning set and the new data as a test set. The TPRs and total classification accuracies with the new data are shown in Table 7.

The highest total classification accuracies were gained by the 5-nearest neighbour method. It classifies 75.0% of

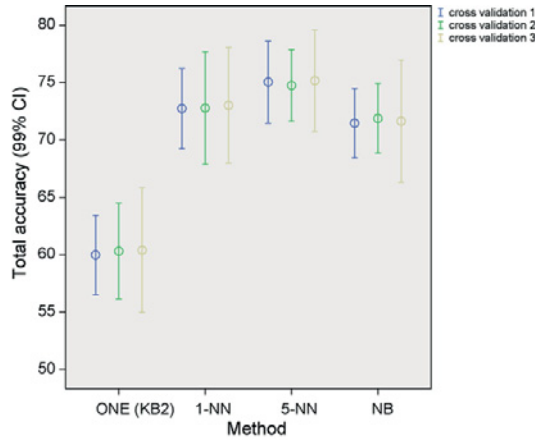


Fig. 5 – Error bars (99% confidence intervals) for total accuracies of each cross-validation with different machine learning methods. The results were obtained with augmented data.

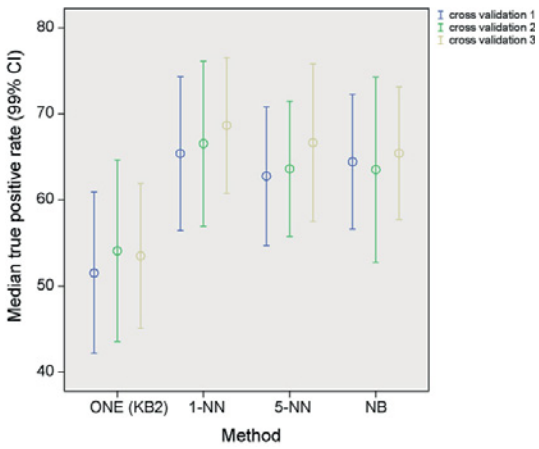


Fig. 6 – Error bars (99% confidence intervals) for the median true positive rates of each cross-validation with different machine learning methods. The results were obtained with augmented data.

the cases correctly in the augmented data and 51.0% in the new data. The 1-nearest neighbour (1-NN) method has better ACC (72.8%) than ONE (60.2%) in the augmented data whereas ONE is better (45.5%) than 1-NN (38.3%) in the new data. In comparison to ONE, the NB classifier yields better accuracies in both data (71.7% and 46.6% in the augmented and new data, respectively). On the basis of the error bars of Fig. 5, the accuracy of ONE is significantly lower than the accuracies of other methods in the augmented data.

In the augmented data set, ONE and 5-NN have the highest TPRs for three classes, 1-NN for two classes and NB for one class. Cases of BPV seem to be especially difficult for ONE. The nearest neighbour methods recognize over 64% of the BPV cases and NB 78.9% of the cases whereas ONE recognizes only 27.4% of the cases as the first diagnosis suggestion. BRV is tricky for all the three methods, only 6.7–15.0% of cases are classified correctly. Vestibulopatia is difficult for the NB classifier. Central lesion is recognized especially weakly with the nearest neighbour methods. The median TPRs of the four methods vary from 61.8% (ONE) to 64.9% (5-NN). No significant differences in the median TPRs were found when considering the error bars of Fig. 6.

With the new data, each method has the highest TPR for one of the four diseases: ONE for vestibulopatia, 1-NN for vestibular neuritis, 5-NN for Ménière's disease, and NB for BPV. BPV cases are difficult for all the methods: at best only 23.8% of the BPV cases are classified correctly. ONE and 1-NN share the best median TPR of 27.5%. For NB and 5-NN the corresponding medians are 26.9% and 23.2%, respectively. According to McNemar's test (with the Bonferroni corrected significance level 0.008) there are significant differences only in the classifications of NN-1 and NN-5.

Table 7 – True positive rates of disease classes and total classification accuracies of ONE for KB2, *k*-nearest neighbour methods and Naive–Bayes classifier in percents in new data

Disease name	Cases	ONE (KB2)	1-NN	5-NN	NB
Benign positional vertigo	80	1.3	12.5	16.3	23.8
Ménière's disease	128	78.9	58.6	82.8	68.8
Vestibular neuritis	20	15.0	35.0	30.0	30.0
Vestibulopatia	25	40.0	20.0	16.0	20.0
Median of TPR		27.5	27.5	23.2	26.9
Total accuracy	253	45.5	38.3	51.0	46.6

6. Discussion

Recognition of different otoneurological disorders can be difficult even for specialists because these disorders have similar kind of symptoms that can vary in different phases of the disease, for example, the beginning of the Ménière's disease can resemble BPV and vice versa. The reason for vertigo can be, for instance, abnormalities in vestibular organ, tension neck, low blood pressure or even tumour [3,38]. Sometimes clinical tests are needed in order to discover the distinguishable signs or reasons for symptoms. Many general practitioners do not encounter frequently patients with vertigo and, therefore, they do not have routine in distinguishing diseases [38]. These practitioners regard all systems supporting diagnostics very useful [38]. The decision support system ONE can be a big help for these physicians: they can use ONE as an assisting tool that shows them questions to ask from a patient and tests needed to be done [3]. It gives also diagnosis suggestions based on the information given about the patient. Currently, ONE is used as an educational tool for medical students at Helsinki University Hospital. It demonstrates diagnostics of vertiginous patients and teaches characteristics of otoneurological disorders, for example, symptoms and clinical tests that are relevant in diagnosing [3,5]. ONE is also used by specialists as a data gathering system for diseases involving vertigo and as a tool supporting decision making with demanding vertigo cases.

Even though the previous studies dealing with ONE have shown its utility, they have revealed topics for further study. In this study, effects of weights on classification performance were a target of our special attention. Further, we wanted to compare ONE's inference to other ML methods.

The results of this study show that there is a need for aid in knowledge discovery. With the experts' knowledge it is possible to get good results but in some classes cases are recognized quite weakly. The reason can be improper weight and/or fitness values set for attributes. With the knowledge discovery method, it is possible to set appropriate fitness values for attributes. However, the fitness values are not adequate enough in separating classes from each other. Particularly, BPV and BRV do get confused with other diseases without weight values. Combining the calculated fitness values with the weight values set by the experts does help classification in certain classes. Still, there are problems especially with the weight values of acoustic neurinoma and Ménière's disease. These classes are recognized better with the weight values

1 and the fitness values calculated from data than using the attribute weighting set by the experts. The results show that there is a need for further development of the knowledge discovery method. It should contain a method for forming preliminary weight values for attributes besides the calculation of fitness values. For weight setting we will test a scattering method [39,40] and methods of feature selection [41,42]. The scattering method was originally developed for studying location of classes in an attribute space and effects of single attributes on classifications.

When the cases of the new data are classified on the knowledge learnt from the augmented data, the classification accuracies are notably weaker than the cross-validation results of the augmented data. The augmented data and the new data were collected with slightly different questionnaires at different institutes in Finland at different times. The difference between these two questionnaires is a consequence of the enhancement of the query base of ONE during 2003–2004. Even though the augmented data was altered to correspond the refined query base, there still exist significant differences in distributions of these two data sets. One reason for the differences can be the data collection with slightly different questionnaires, other the collection at different institutes. Moreover, otoneurological cases are often complex and finding the right diagnosis can be challenging. As it is remarked in [36], medical diagnosis is subjective and, therefore, it can differ significantly depending on the physician doing it and even with the same person at different times.

None of the ML methods can be said to be superior to others: different methods work better for different data [35]. This can be seen also from the results of this study. None of the compared methods, the inference method of ONE, the NB classifier or the 1- or 5-nearest neighbour methods was remarkably superior to others when considering different diseases in both data sets. In the future, our aim is to study more thoroughly classification capabilities of these methods in order to create a hybrid method combining strengths of these methods.

In future research, we will develop a knowledge acquisition tool that can create domain knowledge from domain data and, thus, create a preliminary knowledge base for a new system by using the knowledge discovery method of this study. In order to take advantage of the experts' knowledge with the knowledge discovery method, we are going to make a graphical interface for the knowledge acquisition tool. The graphical interface allows the visualization of the knowledge and makes it easier for experts to make required alterations to knowledge patterns. Combining ONE's expert system shell with this knowledge acquisition tool makes it easier to take the decision support system in use in new institutes and even in totally new domains. Future work will also focus on testing the presented knowledge discovery method and inference mechanism with data from different domains in order to study applicability of the system for various fields.

Conflict of interest

There are no conflicts of interest.

Acknowledgements

The authors wish to thank Erna Kentala, M.D., and prof. Ilmari Pyykkö, M.D., for their aid to the study. The first author acknowledges the support of the Academy of Finland (grants 78676, 104791 and 202185), the University of Tampere Foundation, The Scientific Foundation of the City of Tampere and Finnish Cultural Foundation, Päijät-Häme Regional fund.

The first author dedicates this paper to the memory of her brother Mika (1978–2006).

REFERENCES

[1] M. Havia, E. Kentala, I. Pyykkö, Prevalence of Menière's disease in general population of Southern Finland, Otolaryngol. Head Neck Surg. 133 (5) (2005) 762–768.
[2] H.K. Neuhauser, M. von Brevern, A. Radtke, F. Lezius, M. Feldmann, T. Ziese, T. Lempert, Epidemiology of vestibular vertigo: a neurotologic survey of the general population, Neurology 65 (6) (2005) 898–904.
[3] E. Kentala, A Neurotologic Expert System for Vertigo and Characteristics of Six Otologic Diseases Involving Vertigo, Academic Dissertation, Department of Otorhinolaryngology, University of Helsinki, Finland, 1996.
[4] M. Havia, Menière's Disease Prevalence and Clinical Picture, Academic Dissertation, Department of Otorhinolaryngology, University of Helsinki, Finland, 2004.
[5] Y. Auramo, Construction of An Expert System to Support Otoneurological Vertigo Diagnosis, Academic Dissertation, Department of Computer Sciences, University of Tampere, Finland, 1999.
[6] Y. Auramo, M. Juhola, I. Pyykkö, An expert system for the computer-aided diagnosis of dizziness and vertigo, Med. Inform. 18 (1993) 293–305.
[7] K. Viikki, M. Juhola, Refining the knowledge base of an otoneurological expert system, in: J. Crespo, V. Maojo, F. Martin (Eds.), Medical Data Analysis, Lecture Notes in Computer Science, vol. 2199, Springer, Berlin, 2001, pp. 276–281.
[8] K. Viikki, Machine Learning on Otoneurological Data: Decision Trees for Vertigo Diseases, Academic Dissertation, Department of Computer Sciences, University of Tampere, Finland, 2002.
[9] K. Varpa, K. Iltanen, M. Juhola, E. Kentala, I. Pyykkö, Refinement of the otoneurological decision support system and its knowledge acquisition process, in: R. Engelbrecht, A. Hasman (Eds.), European Notes in Medical Informatics: Ubiquity: Technologies for Better Health in Aging Societies, Proceedings of MIE2006. The 20th International Congress of the European Federation for Medical Informatics, Maastricht, Netherlands, IOS Press, Amsterdam, Netherlands, 2006, pp. 197–202.
[10] T. Mitchell, Machine Learning, McGraw-Hill, New York, NY, 1997.
[11] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo California, 1993.
[12] Y. Yang, G.I. Webb, A comparative study of discretization methods for Naive–Bayes classifiers, in: Proceedings of the Pacific Rim Knowledge Acquisition Workshop (PKAW) 2002, Tokyo, Japan, 2002, pp. 159–173.
[13] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[14] M. d'Aquin, S. Brachais, J. Lieber, A. Napoli, Decision support and knowledge management in oncology using hierarchical

classification, in: Proceedings of the Symposium on Computerized Guidelines and Protocols—CGP-2004. Studies in Health Technology and Informatics 101, IOS Press, 2004, pp. 16–30.
[15] C. Baumgartner, C. Bohm, D. Baumgartner, Modelling of classification rules on metabolic patterns including machine learning and expert knowledge, J. Biomed. Inf. 38 (2) (2005) 89–98.
[16] Y. Huang, P. McCullagh, N. Black, R. Harper, Feature selection and classification model construction on type 2 diabetic patients' data, Artif. Intell. Med. 41 (3) (2007) 251–262.
[17] J. Quentin-Trautvetter, P. Devos, A. Duhamel, R. Beuscart, Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France, Stud. Health Technol. Inf. 90 (2002) 557–561.
[18] R. Shankle, S. Mani, M. Dick, M. Pazzani, Simple models for estimating dementia severity using machine learning, in: Proceedings of MedInfo'98: 9th World Congress on Medical Informatics, Seoul, Korea, 1998, pp. 472–476.
[19] B. Sissons, W.A. Gray, A. Bater, D. Morrey, Using artificial intelligence to bring evidence-based medicine a step closer to making the individual difference, Med. Inform. Internet Med. 32 (1) (2007) 11–18.
[20] N.N. Karanikolas, C. Skourlas, Computer assisted information resources navigation, Med. Inform. Internet Med. 25 (2) (2000) 11–18.
[21] S. Kim, Protein β-turn prediction using nearest neighbour method, Bioinformatics 20 (2004) 40.
[22] T.K. Paul, H. Iba, Extraction of informative genes from microarray data, in: H.-G. Beyer (Ed.), Proceedings of the 2005 conference on Genetic and evolutionary computation, ACM Press, New York, 2005, pp. 453–460.
[23] M.E. Mavroforakis, H.V. Georgiou, N. Dimitropoulos, D. Cavouras, S. Theodoridis, Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers, Artif. Intell. Med. 37 (2) (2006) 145–162.
[24] R. Blanco, P. Larrañaga, I. Inza, B. Sierra, Selection of highly accurate genes for cancer classification by estimation of distribution algorithms, in: Workshop of Bayesian Models in Medicine, (held within 8th Conference on AI in Medicine in Europe), AIME2001, Cascais, Portugal, 2001, pp. 29–34.
[25] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L.C. Showe, M.K. Showe, Combining multi-species genomic data for microRNA identification using a Naive–Bayes classifier, Bioinformatics 22 (11) (2006) 1325–1334.
[26] M. Yousef, S. Jung, A.V. Kossenkov, L.C. Showe, M.K. Showe, Naive–Bayes for microRNA target predictions machine learning for microRNA targets, Bioinformatics 23 (22) (2007) 2987–2992.
[27] R. Blanco, I. Inza, M. Merino, J. Quiroga, P. Larrañaga, Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS, J. Biomed. Inform. 38 (2005) 376–388.
[28] S.R. Bhatikar, C. DeGroff, R.L. Mahajan, A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics, Artif. Intell. Med. 33 (2005) 251–260.
[29] W. Li, Y.P. Tsai, C.L. Chiu, The experimental study of the expert system for diagnosing unbalances by ANN and acoustic signals, J. Sound Vib. 272 (1–2) (2004) 69–83.
[30] M. Juhola, H. Aalto, T. Hirvonen, Using results of eye movement signal analysis in the neural network recognition of otoneurological patients, Comput. Methods Programs Biomed. 86 (2007) 216–226.
[31] S. Ahmad, M.M. Gromiha, NETASA: neural network based prediction of solvent accessibility, Bioinformatics 18 (6) (2002) 819–824.
[32] J.S. Spicker, F. Wikman, M.-L. Lu, C. Cordon-Cardo, C. Workman, T.F. Ørntoft, S. Brunak, S. Knudsen, Neural network predicts the sequence of TP53 gene based on DNA chip, Bioinformatics 18 (2002) 1133–1134.
[33] Y. Auramo, M. Juhola, Modifying an expert system construction to pattern recognition solution, Artif. Intell. Med. 8 (1996) 15–21.
[34] K. Viikki, M. Juhola, E. Kentala, I. Pyykkö, Building training data for decision tree induction in the subspecialty of otoneurology, in: Medical Infobahn for Europe: Telematics in Health Care, Proceedings of the MIE2000 and GMDS2000 Congress -cd, Hannover, Germany, BMIG, Ismaning, Germany, 2000.
[35] D.R. Wilson, T.R. Martinez, Improved heterogenous distance functions, J. Artif. Intell. Res. 6 (1997) 1–34.
[36] I. Kononenko, I. Bratko, M. Kukar, Application of machine learning to medical diagnosis, in: R.S. Michalski, I. Bratko, M. Kubat (Eds.), Machine Learning, Data Mining and Knowledge Discovery: Methods and Applications, Wiley, New York, 1998, pp. 1–23.
[37] B. Cestnik, Estimating probabilities: a crucial task in machine learning, in: Proceedings of the European Conference on Artificial Intelligence (ECAI-90), Stockholm, 1990, pp. 147–149.
[38] P. Aalto, Equihear-markkinaselvitys. Kuulon ja huimauksen IT-pohjainen konsepti, Finn-Medi Tutkimus, Tampere, Finland, 2005.
[39] M. Siermala, M. Juhola, J. Laurikkala, K. Iltanen, E. Kentala, I. Pyykkö, Evaluation and classification of otoneurological data with new data analysis methods based on machine learning, Inf. Sci. 177 (2007) 1963–1976.
[40] M. Siermala, M. Juhola, Techniques for biased data distributions and variable classification with neural networks applied to otoneurological data, Comput. Methods Programs Biomed. 81 (2006) 128–136.
[41] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.
[42] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (1997) 131–156.

PUBLICATION III

Attribute Weighting with Scatter and Instance-Based Learning Methods Evaluated with Otoneurological Data

Kirsi Varpa, Kati Iltanen, Markku Siermala and Martti Juhola

International Journal of Data Science 2(3), 2017, pp. 173–204
https://doi.org/10.1504/IJDS.2017.10007392

Attribute weighting with Scatter and instance-based learning methods evaluated with otoneurological data

Kirsi Varpa, Kati Iltanen, Markku Siermala and Martti Juhola*

Computer Science,
School of Information Sciences,
FI-33014 University of Tampere, Finland
E-mail: kirsi.varpa@gmail.com
E-mail: Kati.Iltanen@uta.fi
E-mail: markku.siermala@gmail.com
E-mail: Martti.Juhola@uta.fi
*Corresponding author

Abstract: Treating all attributes as equally important during classification can have a negative effect on the classification results. An attribute weighting is needed to grade the relevancy and usefulness of the attributes. Machine learning methods were utilised in weighting the attributes. The machine learnt attribute weighting, weights defined by the application area experts, and the weights set to 1 were tested on otoneurological data with the nearest pattern method of the decision support system ONE and the attribute weighted *k*-nearest neighbour method using One-vs-All classifiers. The effects of attribute weighting on the classification performance were examined. The results showed that the extent of the effect the attribute weights had on the classification results depended on the classification method used. The weights computed with the Scatter method improved the total classification accuracy compared with the weights 1 and the expert-defined weights with ONE and the attribute weighted 5-nearest neighbour OVA methods.

Keywords: machine learning; attribute weighting; Scatter attribute importance evaluation method; instance-based learning; attribute weighted *k*-nearest neighbour method.

Reference to this paper should be made as follows: Varpa, K., Iltanen, K., Siermala, M. and Juhola, M. (xxxx) 'Attribute weighting with scatter and instance-based learning methods evaluated with otoneurological data', *Int. J. Data Science*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Kirsi Varpa received her MSc in Computer Science in 2005 at the University of Tampere, Finland. At present, she is a researcher and postgraduate at the University of Tampere. Her research focuses on data analysis, machine learning, knowledge discovery and classification methods.

Kati Iltanen is a lecturer at the University of Tampere, Finland. She received her MSc in 1997 from the University of Kuopio and PhD in 2002 from the University of Tampere. Her research interests include knowledge discovery, data mining and machine learning.

Markku Siermala received his PhD in 2002 in Computer Science at the University of Tampere, Finland. His research interests include topics in machine learning, algorithmics and bioinformatics. He has been working in software industries since 2005.

Martti Juhola received his PhD in 1987 in Computer Science at the University of Turku where he worked in 1980–1992. He was a Professor at the University of Kuopio from 1992 to 1997. Since 1997, he is a Professor at the University of Tampere, Finland. His research consists of topics in biomedical signal analysis, pattern recognition, data analysis and mining.

1 Introduction

Treating all attributes as equally important during classification can have a negative effect on the results by giving noisy, redundant and/or irrelevant attributes a higher influence on the results than they should have. This can, for example, reduce the accuracy of the classification (Lee et al., 2007). With instance-based learning methods, such as the *k*-nearest neighbour method (*k*-NN) (Cover and Hart, 1967), that utilise all available attributes in the distance calculation, the noisy and irrelevant attributes may dominate the results (Wettschereck and Aha, 1995). With equal weighting, the noisy, redundant and/or irrelevant attributes have as much effect on the distance calculations as the relevant ones have. Therefore, the attribute weighting and selection is needed to grade the relevancy and usefulness of the attributes - in some domains even class-dependently.

There are two extremes in the emphasis of classification methods on focusing on relevant attributes: at one extreme there are the methods that use all available attributes in the classification and at the other there are the classification and attribute subset selection methods that explicitly attempt to select relevant attributes and reject the irrelevant and redundant ones (Blum and Langley, 1997). Between these extremes there are attribute weighting methods that aim to achieve good scaling behaviour without explicitly selecting subsets of attributes.

Some of the attributes can be discarded during the data pre-processing based on the abundant missing values or the value being constant with all classes. Statistical and attribute selection methods are needed in order to find irrelevant and redundant attributes. The attribute types occurring in the data set determine which methods to apply. Certain methods can be used only with quantitative attributes, whereas some are suitable only for qualitative attributes.

The attribute selection methods can be organised into three categories depending on how they combine the attribute selection search with the construction of the classification model: filter, wrapper and embedded methods (Blum and Langley, 1997; Kohavi and John, 1997; Saeys et al., 2007). Filter methods are independent of the classification models. They use attribute selection to filter attributes to the classification (Blum and Langley, 1997). Filter methods assess the relevance of the attributes by looking only at the intrinsic properties of the data set. Most filter methods calculate an attribute relevance score based on which attributes with a high scoring are kept and attributes with a low scoring are discarded. A subset of attributes with high relevance scores is given to the classifier. The methods used in attribute filtering are, for example, statistical tests for independence (*e.g.*, X^2 test), measures of association with their significance tests (*e.g.*, Pearson correlation coefficient), information gain, regression and principal component analysis (Blum and Langley, 1997; Saeys et al., 2007).

Wrapper methods wrap the attribute selection around the classification process: the classifier itself is used as part of the function evaluating attribute subsets during the search for a good attribute subset (Blum and Langley, 1997; Kohavi and John, 1997). A search in the space between possible attribute subsets (*e.g.*, forward selection, backward elimination or hill climbing) is defined: various subsets of attributes are generated and evaluated with the classification method (Saeys et al., 2007). The attribute subset with the highest evaluation is chosen as the final subset (Kohavi and John, 1997). Wrapper methods have the ability to take into account of attribute dependencies and the interaction between the data and the classifier. Wrapper methods are utilised with, for instance, nearest neighbour methods and case-based reasoning (Blum and Langley, 1997).

Embedded methods embed the attribute selection within the classifier: the search for an optimal attribute subset is already built into the classifier construction (Saeys et al., 2007). Thus, embedded methods are specific to a given learning algorithm. Examples of embedded methods are decision trees and weighted Naïve Bayes (Blum and Langley, 1997; Saeys et al., 2007).

Heuristic search is a common technique in attribute selection (Blum and Langley, 1997). It is utilised to guide the search for an optimal attribute subset (Saeys et al., 2007), especially with wrapper methods. Heuristic search can be started with an empty attribute subset and continued by successively adding attributes (forward selection) or it can be started with all attributes in the attribute subset and continued by successively removing them (backward elimination). There exist also variations combining forward selection and backward elimination, for instance, stepwise forward–backward selection that adds a given number of attributes into the attribute subset and removes another given number of attributes from the subset in each step (Schulerud and Albregtsen, 2004). Filter, wrapper and embedded methods can be utilised in heuristic search.

The attribute weighting methods can be distinguished in a five-dimensional framework: they can be separated into feedback, weight space, representation, generality and knowledge dimensions (Wettschereck and Aha, 1995). The feedback dimension can be divided by the way the attribute weighting methods assign weights to performance feedback and to ignorant methods. The performance feedback methods, as incremental hill climbers (*e.g.*, the incremental instance-based learning method IB4 (Aha, 1992) and Relief (Kira and Rendell, 1992)) and continuous optimizers (*e.g.*, the genetic algorithm combining its optimisation capabilities with the classification capabilities of the weighted *k*-nearest neighbour algorithm GA-WKNN (Kelly and Davis, 1991)), modify the weights to increase the similarity of case x with the nearby cases of the same class and to reduce the similarity with the cases of the other classes. With the ignorant methods, the attribute weights are assigned with pre-existing models, such as conditional probability, class projection or mutual information (Wettschereck and Aha, 1995). The weight space dimension defines the size of the search space of the weights and differentiates attribute selection from attribute weighting methods; during attribute selection the search space is usually constrained to binary values (0 or 1), whereas attribute weighting uses continuous values (Wettschereck et al., 1997). In the representation dimension, the methods are distinguished by the way they handle an attribute set: is the set used as it was given or is it transformed before weighting. The generality dimension divides the methods into global and local weight setting methods. In global setting it is assumed that a single weight set can describe the whole domain, whereas in local setting the weights can differ among the values of the attributes and even be case-specific (Wettschereck and Aha, 1995). The knowledge dimension separates the attribute weighting methods into knowledge-poor and knowledge-intensive methods, depending on how they employ domain-specific

knowledge in the weighting. The above-mentioned dimensions with different weighting methods are explored in more detail in Wettschereck et al. (1997).

Machine learning (ML) and statistical methods have been utilised in setting weights for attributes needed in other machine learning methods. For example, the properties of a decision tree have been applied to set weights to the Naïve Bayes classifier (the minimum depth of the attribute) (Hall, 2007) and the *k*-nearest neighbour method (path-specific information gain) (Cardie and Howe, 1997), the attribute weights for the *k*-nearest neighbour method have been calculated with a genetic algorithm (Kelly and Davis, 1991; Lee et al., 2007) and from a score based on the X^2 test statistic (Vivencio et al., 2007) and neural network (Zeng and Martinez, 2004) (strength of related links in the neural network), and weights for the attributes have been computed from a collaborative social network using regression analysis (Debnath et al., 2008). Also, the perceptron updating rule can be considered an attribute weighting method in addition to the least-mean squares algorithm and the back-propagation method (Blum and Langley, 1997). Filter, wrapper and embedded approaches have been applied in attribute weighting: the X^2 statistical test is a filter method (Vivencio et al., 2007), IB4 (Aha, 1992) is an embedded method and the genetic algorithm is a wrapper method (Kelly and Davis, 1991).

We are interested in the onward development of an otoneurological decision support system ONE (Auramo et al., 1993) that supports the diagnostics of vertigo diseases. Diagnosis of the otoneurological disorders is demanding because the diseases can simulate each other with symptoms of a similar kind and the symptoms can vary over time, making recognition difficult (Havia, 2004; Kentala, 1996). The system gives diagnosis suggestions for new cases with an inference method utilizing the class-wise weights and fitness values given to the attributes and their values in a knowledge base. Each attribute refers to a sign, a symptom or a measurement data from a clinical test (Auramo et al., 1993). The attribute value indicates, for example, whether the patient has a hearing loss (yes/no), how long the vertigo attacks last (no attacks, less than 1 min, 1 min to 20 min, 20 min to 4 h, 4 to 24 h or more than 1 day) or what the audiometry value is at 2000 Hz (-10–140 dB). The attribute weights and fitness values of the attribute values describe the symptoms, signs and measurement results related to the class; the attribute weight expresses the significance of the attribute for the class, whereas the fitness value describes which attribute values fit the class.

An earlier study showed the need for further enhancement of the knowledge discovery method of ONE (Varpa et al., 2008). Previously, the fitness values for the attribute values were computed by a machine learning method, but all the attributes were equally weighted (each attribute had the weight 1). This alone enhanced the classification accuracy compared with the knowledge descriptions defined purely by the domain experts, but there were still difficulties in the recognition of certain disease classes. The attribute weights defined by the experts were tested with the machine-learnt fitness values, but this combination did not improve the classification as hoped. Therefore, in this study, machine learning methods for attribute weight calculation are applied in order to improve the classification of vertigo diseases.

The methods used for attribute weighting in this research are the Scatter method for attribute importance evaluation (Juhola and Siermala, 2012; Siermala et al., 2007) and the weight calculation method of the incremental instance-based learning algorithm IB4 (Aha, 1992). These methods were selected because they can express the relevance of a single attribute and can learn attribute weights separately for each class. The Scatter method does not have any prerequisites for the class distributions (Juhola and Siermala, 2012). It can be used in attribute filtering, for example, by applying the scatter values in attribute weighting or in the attribute subset selection. The Scatter method is based on

traversing through a data set by seeking the nearest case one at a time and concurrently counting the class changes between cases. A scatter value expresses the attributes' power to separate classes in the data set (Juhola and Siermala, 2012). In this study, the scatter values are calculated for each attribute in a different class versus other classes' situations. The results of the Scatter method were promising in earlier studies (Juhola and Siermala, 2012), so, it was used in this study. The weight calculation method of the IB4 classification method computes attribute weights independently for each class with a simple performance feedback algorithm (Aha, 1992). The attribute weights of IB4 reflect the relative relevancies of the attributes in the class. The methods are described in more detail in section 3.2. The Scatter and IB4 methods both use a continuous weight space and a given representation, calculate local weight settings and do not employ specific domain knowledge in attribute weight setting. They both use pure data in weight setting. Scatter and IB4 differ in the way they handle feedback: IB4 is a performance feedback method that alters the weights based on the classification results during processing, whereas Scatter creates weights based on the pre-existing model and ignores the classification results during the runs.

Machine-learnt attribute weights are utilised with the inference mechanism of the otoneurological decision support system ONE and with the attribute weighted *k*-nearest neighbour method (*wk*-NN) (Kelly and Davis, 1991; Mitchell, 1997) using One-vs-All (OVA) classifiers (Rifkin and Klautau, 2004). Otoneurology is a difficult domain by itself, and with small disease classes and classes containing cases with confounding symptoms included in the data classification of the vertigo diseases it is even more challenging. Therefore, it is good to test the attribute weights with two machine learning methods that have different approaches to the classification: with ONE, that searches for the most compatible class pattern for the case, and with the attribute weighted *k*-NN OVA, which classifies cases based on their nearest instances. The selected methods resemble each other in the way they handle classes separately. The classification accuracies yielded by the different attribute weight and fitness value combinations are compared with each other and with the accuracies of the knowledge formed purely by the experts. In addition, the pair-wise agreement between the machine and human expert classifications is examined using Cohen's kappa (Cohen, 1960).

2 Material

In this study, otoneurological data having 1030 cases from nine different vertigo diseases (classes) was used (Table 1). The data was collected over a decade starting from the 1990s in the Department of Otorhinolaryngology at Helsinki University Central Hospital, Finland, where experienced specialists confirmed all the diagnoses. The class distribution of the data is imbalanced: over one-third of the cases belong to the Menière's disease class, whereas the smallest groups have only around 2 % of the cases.

The data set includes 176 attributes concerning a patient's health status: occurring symptoms, medical history and clinical findings in otoneurologic, audiologic and imaging tests (Kentala et al., 1995; Viikki, 2002), from which 38 attributes are central (Siermala et al., 2007). Clinical tests were not done for each patient and the values of the attributes are missing in several test results. Attributes with low frequencies of available values were not used in this research. After leaving out the attributes having over 35% missing values, 94 attributes remained to be used in this research: 17 quantitative (integer or real value) and 77 qualitative attributes (of which 54 were binary (yes/no), 20 were ordinal and 3 nominal). Almost half of the remaining 94 attributes (46) have less than 5%

missing values and 73 (77.7%) have less than 10% missing values. Only one attribute has information from all cases. Thirteen attributes, all concerning clinical findings, have over 29% of their values missing, and for one important attribute (type of hearing loss) even 53% of the values were missing. The type of hearing loss is crucial in the recognition of sudden deafness and could not be excluded from the data set.

Table 1 The frequency distribution of vertigo disease classes

	Disease name	Abbreviation	Frequency	%
1	Acoustic neurinoma	ANE	131	12.7
2	Benign positional vertigo	BPV	173	16.8
3	Menière's disease	MEN	350	34.0
4	Sudden deafness	SUD	47	4.6
5	Traumatic vertigo	TRA	73	7.1
6	Vestibular neuritis	VNE	157	15.2
7	Benign recurrent vertigo	BRV	20	1.9
8	Vestibulopatia	VES	55	5.3
9	Central lesion	CL	24	2.3
	Total		1030	100

The original data with missing attribute values was used in the classification runs of ONE and the attribute weighted *k*-nearest neighbour method, and in the fitness value computation. It was necessary to impute the data for the attribute weight computation because the Scatter method needs complete input data to work properly. The IB4 method can handle missing attribute values, but, in order to keep it comparable with the Scatter method, the imputed data was also used in its weight calculation. If only the complete cases in the original data had been used, the training set would have been too small. With 94 attributes, there were only 22 complete cases (2.1 %). The number of missing attribute values (9.8 %) allowed the use of imputation. The imputation was done class-wise on the basis of the whole data prior to data division into training and testing sets. The missing values of the attributes were imputed (substituted) with the class modes of the qualitative and the class medians of the quantitative attributes. These simple imputation methods have been proven to be adequate enough for this otoneurological data (Laurikkala et al., 2000).

3 Methods

3.1 Weight utilizing methods

3.1.1 Nearest pattern method of ONE

The inference mechanism of the otoneurological decision support system ONE resembles the nearest neighbour methods of pattern recognition (Auramo and Juhola, 1996). Instead of looking for the nearest case, it looks for the most fitting class for a new case in its knowledge base. In the knowledge base of ONE, a pattern is given to each class that corresponds to one vertigo disease. The pattern can be considered a profile of a disease as it describes its related symptoms and signs. Confounding symptoms are also

acknowledged in the pattern, such as age-related hearing loss and other symptoms not usually related to the disease.

Each class in the knowledge base is described with a set of attributes with weight values expressing their significance for the class. In addition, a fitness value for each attribute value is given to describe how it fits the class (Figure 1).

(a) <attribute name> <attribute weight> <attribute type>
<minimum value> <maximum value>
<value 1> < fitness value 1>

…

<value *n*> < fitness value *n*>
END
.

(b) ATT_OFTEN 4 V
0.0 5.0
0.0 0.0
1.0 1.12
2.0 23.60
3.0 19.10
4.0 43.82
5.0 100.0
END
.

Figure 1 (a) The general form of an attribute pattern in the knowledge base of ONE and (b) an example attribute description ATT_OFTEN (frequency of vertigo attacks with benign positional vertigo)

The weight values vary from 0 to a chosen maximum, where 0 means that the attribute does not concern the class at all. The greater the weight value, the more important the attribute is for the class. Fitness values can have values between 0 and 100. The fitness value 0 means that the attribute value does not fit the class, whereas the fitness value 100 shows that the value fits the class perfectly.

The inference mechanism of ONE (Auramo and Juhola, 1996) searches for the best fitting class in its knowledge base. It calculates scores for the classes from the weight and fitness values of the attributes. The score *S(c)* for a class *c* is calculated in the following way

$$S(c) = \frac{\sum_{a=1}^{A(c)} x(a)w(c,a)f(c,a,j)}{\sum_{a=1}^{A(c)} x(a)w(c,a)}, \qquad (1)$$

where *A(c)* is the number of the attributes associated with the class *c*,
x(a) is 1 if the value of attribute *a* is known and otherwise 0,
w(c,a) is the weight of the attribute *a* for the class *c* and
f(c,a,j) is the fitness value for the value *j* of the attribute *a* for the class *c*

(Auramo and Juhola, 1996). In the case of quantitative attributes, the fitness values are interpolated by using the attribute values in the knowledge base as interpolation points. The fitness values are altered to the range of 0 to 1 during the inference process. The class pattern having the highest score is the best diagnosis suggestion.

In order to handle uncertainty caused by the missing attribute values, ONE calculates the minimum and maximum scores for the classes using the lowest and the highest fitness values for the attributes having missing values. The closer the minimum and maximum scores are to each other, the more reliable the inference result is. There can be diagnosis suggestions having exactly the same highest score (and minimum and maximum score and their difference). In that case, the order of the suggestions having the same score is randomized and the first class is randomly selected from the tied diagnosis suggestions.

3.1.2 Attribute weighted *k*-nearest neighbour method with One-vs-All classifiers

The other method utilizing the weighting schemes is the attribute weighted *k*-nearest neighbour method with One-vs-All classifiers (*wk*-NN OVA). The distance measure of the basic *k*-nearest neighbour method (Cover and Hart, 1967) was expanded to take the attribute weighting into account (Kelly and Davis, 1991; Mitchell, 1997). In addition, in order to keep ONE and the *k*-nearest neighbour method comparable, we decided to convert the multi-class classification problem into multiple binary classifiers - *i.e.*, to divide the *m* class problem into *m* binary problems by using One-vs-All classifiers with the *k*-nearest neighbour method (Galar et al., 2011). Thus, the OVA classifiers and ONE both handle class-wise information, from which the class of a new case is predicted. Each binary OVA classifier was trained to separate a class from all the other classes by marking the cases of this one class as member cases and the cases of the other classes as non-member cases in the training set.

The attribute weighted *k*-NN OVA is an instance-based learning method that searches for the *k* most similar cases (neighbours) of a new case from each classifier separately. There is one classifier per each class and each classifier gives a vote for the case being a member or non-member of the class based on the majority class of the *k* neighbours. The final class of the new case is assigned from a classifier suggesting the case being a member of a class. There can be a situation in which the new case gets more than one member of a class vote (a tie situation) or all of the classifiers vote for the other class (the case to be a non-member of all the classes). In a tie situation, the class of the new case is determined by searching for the most similar member case from the member voting classifiers. The case gets the class of the member case with the shortest distance to it. When all the classifiers vote for the case to be a non-member, the basic attribute weighted 1-nearest neighbour classifier using the whole training data containing the original disease classes is employed to find the most similar case (and its class) for the new case.

The similarity between the new case and the training cases within the classifiers is calculated with a distance measure. In this study, the distance measure used in the attribute weighted *k*-nearest neighbour method was the Heterogeneous Value Difference Metric (HVDM) (Wilson and Martinez, 1997) with attribute weighting, which can handle both qualitative and quantitative attributes in the data set. The attribute weighted HVDM is defined as

$$weighted\_HVDM\,(x,y) = \sqrt{\sum_{a=1}^{m} w_{c_a} d_a(x_a, y_a)^2}\,, \tag{2}$$

where m is the number of attributes,
w_{c_a} is the weight of the attribute a in class c and
$d_a(x_a,y_a)$ is the distance between the values x_a and y_a for attribute a.

The distance function $d_a(x_a,y_a)$ is defined as

$$d_a(x_a, y_a) = \begin{cases} 1, & \text{if } x \text{ or } y \text{ is unknown} \\ normalized\_vdm_a(x_a, y_a), & \text{if } a \text{ is qualitative} \\ normalized\_diff_a(x_a, y_a), & \text{otherwise} \end{cases} \tag{3}$$

Because HVDM computes distances to qualitative and other attributes with different measurement ranges, it is necessary to scale their results into approximately the same range in order to give each attribute a similar influence on the overall distance. Thus, the

measurements are normalized (Wilson and Martinez, 1997). The normalized distance to a quantitative attribute is calculated with Equation 4

$$normalized\_diff_a(x_a, y_a) = \frac{|x_a - y_a|}{4\sigma_a}, \quad (4)$$

where σ_a is the standard deviation of the numeric values of attribute a in the training set of the current classifier, and to a qualitative attribute with Equation 5

$$normalized\_vdm_a(x_a, y_a) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2}, \quad (5)$$

where C is the number of output classes in the problem domain (in this case C=2: the data in the training set T of the classifier is divided into the member and non-member classes),
$N_{a,x(y),c}$ is the number of cases in T that have a value x (or a value y) for attribute a and the output class c, and
$N_{a,x(y)}$ is the number of cases in T that have a value x (or a value y) for attribute a

(Wilson and Martinez, 1997). In other words, we are calculating the conditional probabilities to have the output class c when having attribute a with the value x (or the value y).

3.2 Attribute weight setting methods

3.2.1 Domain experts

The original attribute weights and fitness values of attribute values for the knowledge base of the decision support system ONE was defined by a group of experienced otoneurological physicians in the 1990s (Kentala et al., 1998). A decade later, the knowledge base of ONE was updated by two specialists during the upgrade process of the decision support system (Varpa et al., 2006), when new attributes were added into its knowledge base. The fitness values were not yet in use with the binary type attributes then, and, therefore, the experts did not define fitness values for them.

The original weighting was done on the basis of the experts' knowledge and experience, and on information obtained from the medical literature and data (Kentala et al., 1998). For example, the newest diagnostic criteria for diseases were obtained from medical journals. Furthermore, the collected data on several hundred patients was employed in the knowledge formation (Auramo and Juhola, 1995). The original knowledge base was used as a starting point in the knowledge updating process. The experts went through the attributes and weights in the disease patterns one by one and changed the weighting if necessary (Varpa et al., 2006). Weights were manually defined for each attribute in each disease pattern. The experts used their knowledge and experience as the basis when defining the weights. In addition, they were able to compare their assumptions about the diseases with the collected data (all 1,030 cases) during the updating process.

The medical experts could define weights and fitness values for seven disease classes: acoustic neurinoma, benign positional vertigo, Menière's disease, sudden deafness, traumatic vertigo, vestibular neuritis and benign recurrent vertigo. Two classes (vestibulopatia and central lesion) were found to be too complex to describe with weight

and fitness values. Therefore, in classification runs with the experts' knowledge, seven disease classes with 951 cases were used in this study.

3.2.2 Scatter method

The first machine learning method applied to the attribute weight setting is the Scatter method (Juhola and Siermala, 2012; Siermala et al., 2007). The Scatter method can be utilised to evaluate whether a data set includes meaningful information that can be used for class separation. It has been used, for example, to solve the importance and separation power of attributes and to map the overlap of the classes in the attribute space. A scatter value describes the power of an individual attribute or attribute set to separate the classes in the data. In this study, we were interested in each attribute's power to differentiate one class from the other classes and the possibility to transform the scatter values into weights that can be utilised in the classification. Therefore, scatter values were separately computed for all attributes within each disease class vs. all the other classes.

In order to calculate the scatter value, the entire data set must be traversed through from a case to its nearest unvisited neighbour case. Before calculation, the attribute values are normalized into the same scale [0, 1]. The Scatter method starts by randomly selecting an initial case x from the data. The nearest case y for x is searched with the Euclidean distance. If there are several cases with exactly the same distance, the nearest case y is randomly selected from these nearest cases. The classes of x and y are compared: If the cases are from different classes, a counter a is incremented; otherwise, (they are from the same class) a is kept unchanged. After the comparison, case x is removed from the data set and case y is set as a new x. A new nearest case y is searched from the diminished data set and the classes are compared. These steps are repeated until only case x is left in the data set. After going through all the cases in the data set, the scatter value s is calculated with Equation 6

$$s = \frac{a}{A}, \tag{6}$$

where a is the total number of observed changes between the classes and
A is the theoretical maximum number of possible class changes.

A is computed as follows (Equation 7): Let m_G be the size of the largest class and M_O be the sum of the number of cases in the other classes (in other words, M_O = n- m_G, where n is the number of cases in the data set). When m_G is greater than M_O, A is equal to $2M_O$ and, otherwise, A equals n-1.

$$A = \begin{cases} 2M_{\mathrm{O}}, & \text{if } m_{\mathrm{G}} > M_{\mathrm{O}} \\ n-1, & \text{if } m_{\mathrm{G}} \leq M_{\mathrm{O}} \end{cases} \tag{7}$$

Thus, the scatter value s describes the relationship between the number of observed class changes and the theoretical maximum number of changes. The scatter values vary in (0, 1]. The closer the scatter value is to 0, the more accurately separated from each other the classes are in the attribute space. The scatter value is close to 1 if the cases are selected alternately from different classes, meaning that the classes are entirely overlapping in the attribute space. The Scatter method is described in more detail in (Juhola and Siermala, 2012).

The scatter value describes the overlap of the classes within the attribute values: the closer the scatter value is to 0, the better the attribute differentiates the classes. Nevertheless, the interpretation of the attribute weight values is opposite to the scatter

values: the greater the weight value, the more important the attribute is. Therefore, we needed to take inverses of the scatter values in order to use them as attribute weights.

3.2.3 Instance-based learning algorithms IB4 and IB1w

The other machine learning method applied to the attribute weight formation is Aha's attribute weight learning algorithm from the incremental instance-based learning algorithm IB4 (Aha, 1992). IB4 tolerates irrelevant attributes by learning attribute relevancies (*i.e.* weights for attributes) independently for each class and using these weights in its similarity function. It can also handle skewed class distributions. The learnt attribute weights are receiving our special attention and we do not report the classification results of IB4, but we do use the learnt weights with the nearest pattern method of ONE and the attribute weighted *k*-nearest neighbour OVA method.

In the IB4 method, each class c is described with a separate class description CD_c and a set of attribute weights $Weight_{c_a}$ (Aha, 1992). The class description contains a set of cases with classification records about their past performance during classification, that is, their number of correct and incorrect classification predictions. Based on their classification performance, the cases stored in CD_c are defined as statistically acceptable or mediocre. Cases in CD_c are regarded as statistically acceptable if their classification accuracy is statistically significantly greater than their class's observed frequency (the statistical calculation is based on the confidence intervals) (Aha, 1992; Aha et al., 1991). Acceptable cases are used in the subsequent classification tasks. If there are no acceptable cases in CD_c, mediocre cases are used in the classification instead. Mediocre cases are kept in the class description as long as they are regarded as noisy. Noisy cases with significantly poor classification performance (classification accuracy statistically significantly less than the class's observed frequency) are discarded from the CD_c as soon as they are revealed. The status of the saved cases in CD_c can change during the learning of the attribute weights: mediocre cases can change to noisy or acceptable and even cases previously regarded as acceptable can be discarded from the description when they later appear to be noisy.

In the beginning, a class description is empty and the attribute weights are zero. The first learning case x is moved directly into the class description. When there is at least one case in the class description, the similarity between the learning case x and the cases in CD_c are calculated with an attribute weighted negative Euclidean distance measure

$$Similarity(c, x, y) = -\sqrt{\sum_{a=1}^{m} Weight_{c_a}^2 (x_a - y_a)^2} \qquad (8).$$

The attribute values of x and y are normalized to the range [0, 1] in order to have the same (maximal) effect on the similarity with each attribute. If x_a or y_a is missing, these values are assumed to be maximally different, *i.e.*, the difference $(x_a - y_a)$ is 1. The most similar acceptable neighbour is searched from the CD_c and set as the nearest neighbour y_{max}. If there are several acceptable cases with the same highest similarity, the class frequency within these cases is checked and a case from the class having the highest frequency is randomly selected as y_{max}. If there are no statistically acceptable cases in the CD_c, a random number i is selected within $[1, |CD_c|]$ and the ith most similar case from the CD_c is set as the nearest neighbour y_{max} (Aha et al., 1991). The classes of x and y_{max} are compared. When the classes of x and y_{max} are different (x is misclassified), x is added to the class description CD_c. After the classification of x, the classification records of all saved cases in CD_c that are at least as similar as y_{max} are updated (the number of correct or

incorrect classification predictions are increased, depending on whether or not the class was correct). The saved cases regarded as noisy are discarded from the CD_c. In addition, all attribute weights are adjusted after the classification of each learning case x through a performance feedback algorithm (described in Algorithm 1) to reflect the relative relevancies of the attributes: the weights of attributes are increased when they correctly predict the classification and are otherwise decreased. The attribute weights are defined in the range [0, 0.5], where the weight 0 means that the attribute is irrelevant (Aha, 1992). The weight range is set to [0, 0.5] instead of [0, 1] because the total weight of an irrelevant attribute is expected to be half of its total possible attribute weight (Aha, 1992).

Algorithm 1 The attribute weight updating algorithm of IB4 (Aha, 1992)

Since the cases are normalized, step 1 yields a value in [0,1].

Attributes: x = case being classified
y_{max} = the classifying case from CD_c
c = the target class
λ = the higher observed relative frequency among x's actual and predicted (y_{max}) class members, value range [0,1]

For each attribute a:

1. LET *difference* = $| x_a - y_{\max_a} |$
2. IF (x's classification was correctly predicted ($x\_class == y_{max}\_class$)
   THEN $Cumulative\ Weight_{c_a} = Cumulative\ Weight_{c_a} + (1-\lambda)*(1-difference)$
   ELSE $Cumulative\ Weight_{c_a} = Cumulative\ Weight_{c_a} + (1-\lambda)*difference$
3. $WeightNormalizer_{c_a} = WeightNormalizer_{c_a} + (1-\lambda)$
4. $\text{Weight}_{c_a} = \max\left(\frac{Cumulative\ Weight_{c_a}}{WeightNormalizer_{c_a}} - 0.5, 0\right)$

The novel learning case x is classified in each class description (in this study to seven and nine disease classes). Since the classes are represented separately, the cases are either members or non-members of the class. As a result, there are separate class descriptions and attribute weight sets for each disease class used.

In addition, the attribute weight algorithm was applied with IB1 (Aha et al., 1991), a simpler version of the instance-based learning algorithm IB4. This was done because of the imbalanced class distribution of the data in use: we wanted to see if there were any differences in the attribute weights when handling the class descriptions in different ways. We needed to modify the original IB1 method in order to use it appropriately in this research. First of all, the weighted similarity function (Equation 8) was taken into use with the IB1 method. IB1 usually handles all classes at the same time with one classifier. The weight values are needed for each class separately. Therefore, we needed to alter the IB1 method to work like IB4, having class descriptions for each class separately. This variant of IB1 is called IB1w. The difference between IB1w and IB4 is that IB1w saves all processed cases in its class descriptions and does not discard any cases from the class descriptions during runs. Also, the cases with poor classification records are kept in class descriptions with IB1w.

3.3 Cross-validation

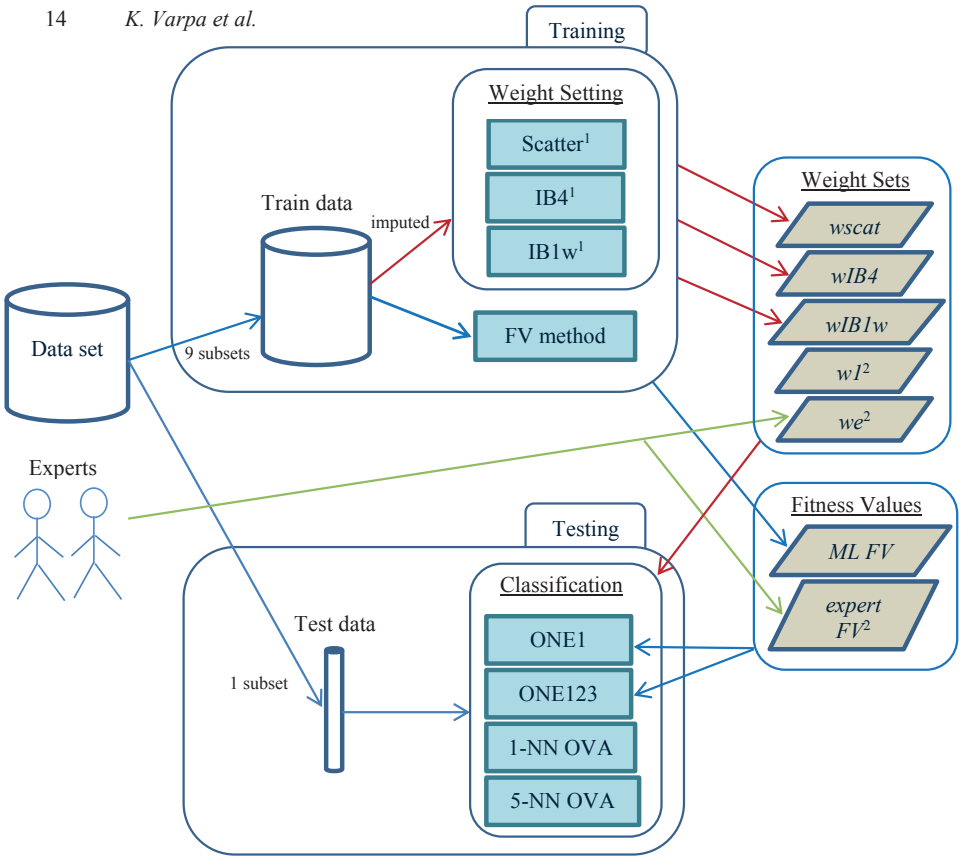
We used a 10-fold cross-validation (CV) (Mitchell, 1997) to evaluate the classification performance of the ONE and the attribute weighted *k*-nearest neighbour OVA methods combined with different weighting schemes. In the 10-fold cross-validation, the data was randomly divided into 10 subsets of approximately equal size. The division was made in a stratified manner to ensure that the class distribution of each subset resembled the skewed class distribution of the entire data set. In the 10 training and testing runs, each training data set included the cases of nine subsets and the testing data set included the cases of the remaining subset. The 10-fold cross-validation was repeated 10 times. Thus, in total, there were 100 runs per each classification method - weighting scheme combination. The same cross-validation divisions were used with all the combinations - *i.e.*, each combination had the same training and testing data sets used during the runs.

The class-wise fitness values (FV) of the attribute values for the nearest pattern method of ONE were computed once for each CV training data set with the fitness value method described in study (Varpa et al., 2008). The original data set containing also the cases having missing attribute values were used in the fitness value calculation. The fitness values for attribute values by experts' were defined only once.

The attribute weights were calculated for each CV training set from the imputed data. The calculation of the weights within each CV training set was repeated 10 times with the Scatter, IB4 and IB1w methods in order to handle the randomness in these methods. In the Scatter method, the randomly selected starting case and, possibly, randomly selected nearest cases when having several neighbours with the same distance both have an effect on the final result of the calculation. In the IB4 and IB1w methods, the order of the cases in the data set affects the results (Wettschereck and Aha, 1995) and, therefore, the order of the cases was mixed up within the repetitions. The mean weights of the 10 weight calculation repetitions were saved into weight sets and used in the classification. Attribute weights were necessary to calculate separately for seven and nine disease class classifications. The attribute weights defined by the application area experts (*we*) were the same in each CV run.

In order to prepare for a possible situation where all classifiers in the attribute weighted *k*-nearest neighbour method with OVA vote a case to be a non-member, it was necessary to calculate the Scatter-based weight values and IB4 and IB1w weights from the training data set having the original classes in addition to the class-wise attribute weights. In the OVA non-member voting situation, the basic attribute weighted *k*-nearest neighbour method with one classifier and one weight set was used. These attribute weight calculations were also repeated 10 times. In the non-member voting situations with the attribute weights defined by the experts we needed to use weights 1 with the basic weighted *k*-NN because the experts could not set a single combination of attribute weights that only contains one weight for each attribute and can separate all disease classes. For the experts, it was more natural to define the class-wise attribute weights by considering the characteristics of a certain disease.

Different weighting scheme and classification method (ONE and attribute weighted *k*-NN OVA) combinations formed for each CV training data set were tested with corresponding CV testing data sets using the original cases with the missing attribute values. The research process in a CV run of the 10-fold CV is summarized in Figure 2.



[1] The weight setting methods were run 10 times in each CV run in order to handle randomness within the methods. The weight sets contained the mean weights of 10 runs.
[2] Weight sets *w1* and *we* and experts' fitness values *expert FV* were created only once.

Figure 2 Description of the research process within a cross-validation (CV) run of the 10-fold CV. The 10-fold CV was repeated 10 times, so, this process was repeated 100 times. In addition, the attribute weights were calculated and tested separately for seven and nine disease classes.

4 Results

When testing the effect of attribute weights on the classification performance, five different weight sets were used:

w1 Equal weighting, all attribute weights set to 1.
we Weights set by the experts. The weights varied from 0 to 15, except the weight 40 of the attribute *hl_type* for sudden deafness. The experts could set weights for seven disease classes.

wscat The weights computed with the Scatter method. The attribute weights were inverse scatter values and varied from 1 to 14.

wIB4 The weights computed with the weight calculation method of Aha's IB4 algorithm. Only the statistically acceptable and mediocre cases were kept in the class descriptions during the weight calculations, and the non-acceptable cases were dropped out. The weights varied from 0 to 0.5.

wIB1w The weights computed with the weight calculation method of Aha's IB4 algorithm, but the case handling was derived from Aha's IB1 method: all of the cases were added to the class descriptions and kept there. The weights varied from 0 to 0.5.

These weight sets were used as attribute weights with the machine learnt fitness values in the knowledge base of ONE and with the attribute weighted *k*-nearest neighbour method having OVA classifiers (*wk*-NN OVA). In addition, classification run of ONE with the knowledge base fully formed by the domain experts (*ONE experts*) was used as the basis in the result comparisons. In this knowledge base, both the attribute weights and the fitness values of attributes were defined by the experts. Expert-set attribute weights (*we*) for seven disease classes were used with both classification methods and, in order to have the results comparable with each other, attribute weight values were computed with the machine learning methods from data containing the seven diseases. The attribute weight sets *wscat, wIB4* and *wIB1w* were also formed from data containing all nine disease classes in order to compare the classification performance between the methods with more classes.

The classification performance of the methods with different attribute weight sets is described with a class-wise true positive rate (TPR) and a total classification accuracy (ACC). TPR is calculated as the percentage of correctly inferred cases in the class:

$$TPR = 100\frac{t_{pos_c}}{n_{cases_c}}\%, \tag{9}$$

where t_{pos_c} is the number of correctly classified cases in the class *c* and

n_{cases_c} is the number of all cases in the class *c*.

The total classification accuracy gives the percentage of all correctly classified cases within the data set:

$$ACC = 100\frac{t_{pos}}{n_{cases}}\%, \tag{10}$$

where t_{pos} is the total number of cases correctly classified in all classes and

n_{cases} is the total number of cases used in the classification.

In addition to the classification rates TPR and ACC, classification method – weight set combinations were examined with Cohen's kappa (*K)* (Ben-David, 2007; Cohen, 1960):

$$K = \frac{P_o - P_c}{1 - P_c}, \tag{11}$$

where P_o is the total agreement probability (*i.e.* accuracy) and

P_c is the probability of predicting the correct class due to chance.

Cohen's kappa was used separately for each classification method – weight set combination to estimate the degree of agreement between their classification results and the actual class labels, and, in addition, to evaluate the pair-wise agreement between the

compared combinations. The value range of kappa is [-1, 1], where -1 means total disagreement (worse than random performance), 0 is a random or majority-based classification and 1 is perfect agreement. Usually, when the kappa value is higher than 0.81, the pair is considered to have almost perfect agreement (Landis and Koch, 1977).

When comparing the classification results of the seven disease classes based on the first diagnosis suggestion of ONE and the attribute weighted 1- and 5-nearest neighbour methods with the OVA classifiers using the different attribute weight combinations in Table 2, it can be seen that the highest total classification accuracy (79.7%), the highest median true positive rate (75.2%) and the highest Cohen's kappa (0.73) were achieved with the Scatter weighted 5-nearest neighbour method (*5-NN OVA wscat)*. The other nearest neighbour methods classified 70.8% to 78.9% of the cases correctly, had a median TPR between 60.6% and 74.3% and Cohen's kappa varying from 0.61 to 0.72, whereas the total classification accuracies of ONE combinations varied from 43.3% to 74.6%, with a median TPR between 47.8% and 69.8% and Cohen's kappa from 0.33 to 0.67. The ONE combination having the highest total accuracy and Cohen's kappa (74.6% and 0.67 respectively) was ONE with the Scatter weights (*ONE1 wscat*). The highest median TPR (69.8%) was achieved with ONE using IB1w weights (*ONE1 wIB1w*). Based on the kappa values, all of the weighted *k*-NN OVA and ONE variants except *ONE1 experts* and *ONE1 we* had a substantial agreement with the actual classes (kappa value over 0.6). Error bars (with 99% confidence intervals) for the mean total accuracies, mean median true positive rates and mean Cohen's kappa of ONE and the attribute weighted 1- and 5-nearest neighbour OVA methods with different weighting schemes achieved within 10 times repeated 10-fold cross-validation are shown in Figure 3.

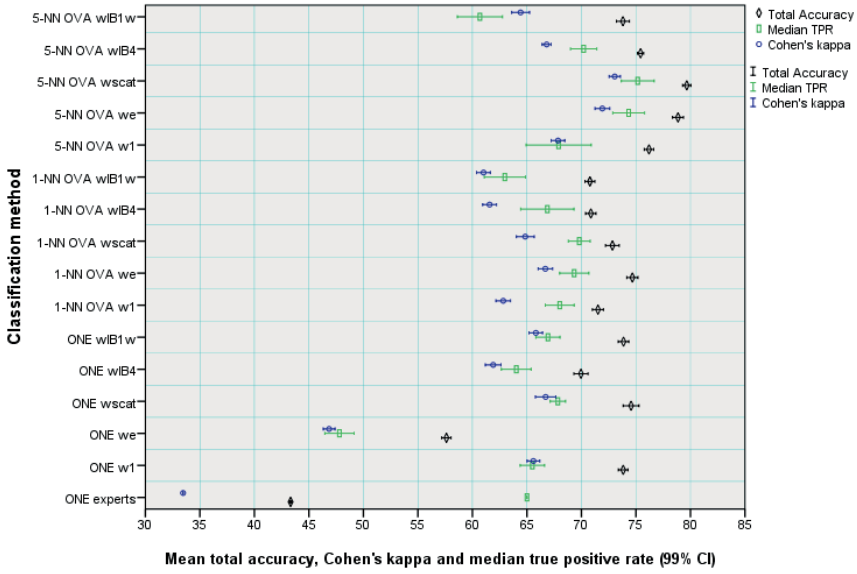


Figure 3 Error bars (with 99% confidence intervals) for the mean total accuracies, Cohen's kappas and median true positive rates (TPR) of classification methods from 10 times repeated 10-fold cross-validation with seven disease classes.

Table 2 The true positive rates (TPR) of seven disease classes and the total classification accuracies with ONE's first diagnosis suggestion (ONE1) and the attribute weighted *k*-nearest neighbour method with OVA classifier (*wk*-NN OVA) in percentages (%) from 10 times repeated 10-fold cross-validation. In addition, the Cohen's Kappa (K) and the Kappa Chance agreement (P_c) are presented. The highest TPRs, accuracy and Kappas are in boldface.

								wk-NN OVA w1		*wk*-NN OVA we		*wk*-NN OVA wscat		*wk*-NN OVA wIB4		*wk*-NN OVA wIB1w	
Disease	Cases	ONE1 experts	ONE1 w1	ONE1 we	ONE1 wscat	ONE1 wIB4	ONE1 wIB1w	1-NN	5-NN	1-NN	5-NN	1-NN	5-NN	1-NN	5-NN	1-NN	5-NN
E	131	24.4	65.6	16.7	62.3	63.8	66.3	**68.5**	64.6	67.6	65.9	63.0	63.1	63.3	60.5	63.0	60.6
BPV	173	65.9	54.7	47.8	55.6	50.3	53.5	69.9	71.8	69.3	**74.3**	69.4	70.9	70.6	70.7	68.4	68.5
MEN	350	42.0	91.7	75.8	91.9	81.4	90.5	82.4	**95.4**	87.2	92.1	80.7	93.7	84.2	94.9	88.5	95.3
SUD	47	68.1	62.6	**85.5**	71.9	68.1	65.1	45.7	29.4	45.3	51.9	68.7	84.3	28.1	25.5	27.4	27.4
TRA	73	67.1	79.0	40.1	83.2	**94.5**	83.7	63.3	67.9	74.8	79.0	80.8	86.6	67.4	72.7	50.5	54.4
VNE	157	15.9	67.8	66.1	67.8	63.8	68.0	68.8	72.8	74.4	**80.7**	70.9	75.2	68.9	73.0	69.2	72.7
BRV	20	**65.0**	36.5	23.5	43.0	43.0	39.5	26.5	20.0	19.0	19.0	25.0	18.5	17.5	19.5	19.5	17.5
Median of TPR		65.0	65.6	47.8	67.8	63.8	69.8	68.5	67.9	69.3	74.3	69.8	**75.2**	67.4	70.7	63.0	60.6
Total ACC	951	43.3	73.8	57.6	74.6	70.0	73.9	71.5	76.2	74.7	78.9	72.8	**79.7**	70.9	75.4	70.8	73.8
K		0.33	0.66	0.47	0.67	0.62	0.66	0.63	0.68	0.67	0.72	0.65	**0.73**	0.62	0.67	0.61	0.64
P_c		**0.15**	0.24	0.20	0.24	0.21	0.24	0.23	0.26	0.24	0.25	0.23	0.25	0.24	0.26	0.25	0.26

In the figure, the Cohen's kappa values are altered to the range [-100, 100] in order to make the figure easier to interpret. The total accuracies, Cohen's kappa and median TPR of two of the weighted 5-NN OVA variants (*5-NN OVA wscat* and *5-NN OVA we*) were significantly higher than the results of the ONE and other *k*-NN OVA variants. *ONE1 wscat, ONE1 wIB4, ONE1 wIB1w* and *ONE1 w1* had similar kind of results with the weighted 1-NN OVA variants. *ONE1 experts* and *ONE1 we* had significantly lower results based on the error bars. The total accuracies and kappa were quite stable between the 10-fold cross-validation runs, whereas the median true positive rates varied by a few percentage points.

The best true positive rates of the disease classes were achieved with different methods and different attribute weights: the highest TPR (95.4%) was achieved on Menière's disease with the 5-nearest neighbour method with weights 1 (*5-NN OVA w1*). The IB4 weighted ONE (*ONE1 wIB4*) had the best TPR for traumatic vertigo (94.5%), ONE with the experts' weights (*ONE1 we*) rated the best sudden deafness (85.5%) cases, the 5-nearest neighbour method with the experts' weights (*5-NN OVA we)* had the highest TPRs for vestibular neuritis (80.7%) and benign positional vertigo (74.3%), the 1-nearest neighbour method with weights 1 (1-*NN OVA w1)* had the best TPR for acoustic neurinoma (68.5%) and ONE purely defined with the experts' knowledge (*ONE1 experts*) had the highest TPR for benign recurrent vertigo (65.0%).

From the classification results of ONE in Table 2 it can be seen that the knowledge bases containing the machine learnt weights (*ONE1 wscat, wIB4* and *wIB1w*) improved the total classification accuracy by more than 26% compared with the knowledge base fully formed by the domain experts (*ONE1 experts*) and more than 12% compared with the knowledge base containing attribute weights defined by the experts (*ONE1 we*): *ONE1 experts* and *ONE1 we* classified 43.3% and 57.6% of the cases correctly and the knowledge bases with the machine learnt weights 74.6%, 70.0% and 73.9% respectively. Knowledge base *ONE1 w1* treating all attributes as equally important performed better than *ONE1 experts* and *ONE1 we* and, in addition, better than *ONE1 wIB4*. Its total classification accuracy was 73.8%.

Interestingly, for the 1-nearest neighbour OVA method, the best total accuracy of 74.7% and Cohen's kappa of 0.67 were achieved with the experts' weights (*1-NN OVA we*), while the second best classifier was 1-*NN OVA wscat* with an accuracy of 72.8% and a kappa value of 0.65. For the 5-nearest neighbour method, the Scatter based weights yielded the best results: *5-NN OVA wscat* classified 79.7% of the cases correctly and had a kappa value of 0.73, whereas *5-NN OVA we* achieved a total accuracy of 78.9% with a kappa value of 0.72 and *5-NN OVA w1* achieved 76.2% and 0.68 respectively. The weights of the IB4 and IB1w methods slightly reduced the total classification accuracy with both 1- and 5-nearest neighbour OVA methods. The best and worst results of the attribute weighted 1- and 5-nearest neighbour methods with OVA classifiers using different weight settings were much closer to each other than the results of ONE.

Classification method – weight set combinations were also evaluated pair-wise with Cohen's kappa within 10 times repeated 10-fold cross-validation runs in order to see the interrelated agreement between two combinations (Figure 4).

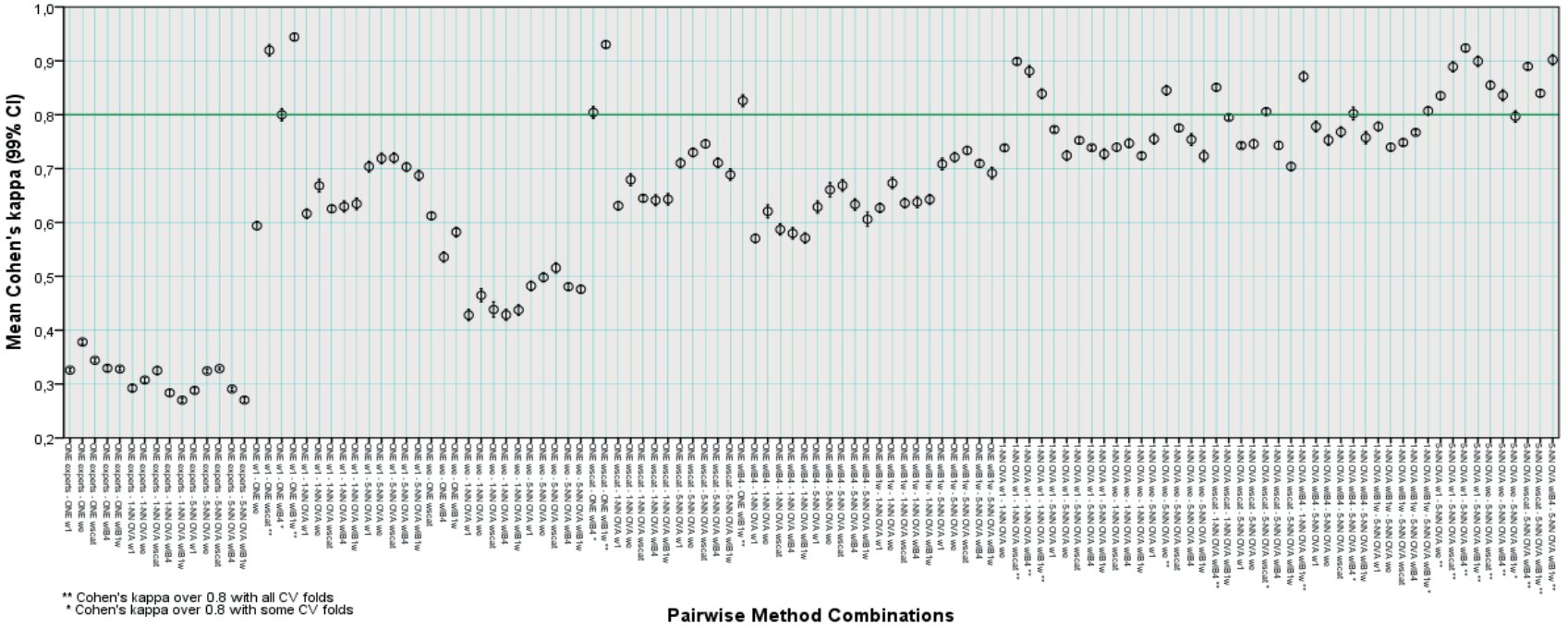


Figure 4 Error bars (with 99% confidence intervals) for the mean Cohen's kappa for pair-wise method combinations from 10 times repeated 10-fold cross-validation with seven disease classes.

There were 19 combination pairs that had almost perfect agreement (a kappa value of over 0.8) on their classification results in every 10-fold run:

ONE1 w1 - ONE1 wscat
ONE1 w1 - ONE1 wIB1w
ONE1 wscat - ONE1 wIB1w
ONE1 wIB4 - ONE1 wIB1w
1-NN OVA w1 - 1-NN OVA wscat
1-NN OVA w1 - 1-NN OVA wIB4
1-NN OVA w1 - 1-NN OVA wIB1w
1-NN OVA we - 5-NN OVA we
1-NN OVA wscat - 1-NN OVA wIB4
1-NN OVA wIB4 - 1-NN OVA wIB1w
5-NN OVA w1 - 5-NN OVA we
5-NN OVA w1 - 5-NN OVA wscat
5-NN OVA w1 - 5-NN OVA wIB4
5-NN OVA w1 - 5-NN OVA wIB1w
5-NN OVA we - 5-NN OVA wscat
5-NN OVA we - 5-NN OVA wIB4
5-NN OVA wscat - 5-NN OVA wIB4
5-NN OVA wscat - 5-NN OVA wIB1w
5-NN OVA wIB4 - 5-NN OVA wIB1w.

Four of these pairs consisted of ONE combinations; the others were 1- and 5- nearest neighbour combinations. In addition to above mentioned pairs, "*1-NN OVA wscat - 5-NN OVA wscat*", "*1-NN OVA wIB1w - 5-NN OVA wIB1w*", "*ONE1 wscat – ONE1 wIB4*", "*1-NN OVA wIB4 - 5-NN OVA wIB4*", "*5-NN OVA we - 5-NN OVA wIB1w*" and *"ONE1 w1 – ONE1 wIB4*" had almost perfect agreement in some of the 10 times repeated 10-fold runs (9, 9, 7, 5, 4 and 3 out of 10 respectively). The Cohen's kappa shows that the 5-nearest neighbour OVA variants with different weight sets are more similar to each other and agree more on the classifications than the 1-nearest neighbour OVA and ONE combinations. Thus, the weight sets have more effect on the classification results of the 1-nearest neighbour OVA and ONE methods than on the results of the attribute weighted 5-nearest neighbour OVA method.

In this domain, patients can have confounding and overlapping symptoms and diseases can mimic other diseases (Havia, 2004; Kentala, 1996), which led us to investigate the number of tied diagnosis suggestions of ONE and tied votes of *k*-NN OVA variants within 10 times repeated 10-fold cross-validation. ONE had only one case with two tied best suggestions (Table 3(a)). With the attribute weighted 1- and 5-nearest neighbour OVA methods, the number of cases having tied voting classifiers was quite large (Table 3(b)). There were situations where the 1- and 5-nearest neighbour method with OVA classifiers voted a case to be a member of more than one classifier or voted it to be a non-member of all classes. The total number of cases having tied voting classifiers varied with the 1-nearest neighbour OVA method from 204 to 279 (from 21.5% to 29.3%) and with the 5-nearest neighbour OVA method from 158 to 228 (from 16.6% to 24.0%) within 10 times repeated 10-fold cross-validation. The lowest total number of cases having tied voting classifiers within 1-NN OVA (from 204 to 223) was yielded with *1-NN OVA wIB4* and within 5-NN OVA (from 158 to 179) with 5-*NN OVA wIB1w*. The proportion of cases having non-member voting classifiers with each 1- and 5- nearest neighbour OVA variant was quite high: at worst, 14.5% of 1-NN OVA and 13.0% of 5-NN OVA cases could not be assigned to a class with the OVA classifiers. In these non-member voting situations, the class was solved using the basic attribute weighted 1-nearest neighbour method.

In order to see what diseases were mixed up with others, we created mean confusion matrices for the classification methods ONE and 1- and 5- nearest neighbour methods using OVA classifiers with the weight combinations that had the highest total accuracy from the 10 times repeated 10-fold cross-validation (Table 4). The confusion matrix of *ONE1 experts* was added for comparison.

Table 3 The minimum and maximum number (*n*) of cases having tied diagnosis suggestions or a tied voting situation occurring in 10 times repeated 10-fold cross-validation with seven disease classes (the number of cases covers the entire 10-fold data).

(a) diagnosis suggestions of ONE with the same highest score and maximum score and minimum score difference.

n of tied voting classifiers	ONE1 experts	ONE1 w1 min *n*	ONE1 w1 max *n*	ONE1 we	ONE1 wscat	ONE1 wIB4	ONE1 wIB1w
2 suggestions	0	0	1	0	0	0	0
total *n* of ties	0	0	1	0	0	0	0

(b) tied voting with the attribute weighted 1- and 5-nearest neighbour methods with OVA classifiers.

n of tied voting classifiers	1-NN OVA w1 min *n*	1-NN OVA w1 max *n*	1-NN OVA we min *n*	1-NN OVA we max *n*	1-NN OVA wscat min *n*	1-NN OVA wscat max *n*	1-NN OVA wIB4 min *n*	1-NN OVA wIB4 max *n*	1-NN OVA wIB1w min *n*	1-NN OVA wIB1w max *n*
2 class members	124	138	115	135	135	150	101	127	125	140
3 class members	2	7	8	14	3	9	2	8	2	7
4 class members	0	0	0	1	0	1	0	0	0	0
5 class members	0	0	0	0	0	0	0	0	0	0
6 class members	0	0	0	0	0	0	0	0	0	0
7 non-members	81	102	124	138	91	106	88	104	73	91
total *n* of ties	216	237	253	279	240	260	204	223	212	232

n of tied voting classifiers	5-NN OVA w1 min *n*	5-NN OVA w1 max *n*	5-NN OVA we min *n*	5-NN OVA we max *n*	5-NN OVA wscat min *n*	5-NN OVA wscat max *n*	5-NN OVA wIB4 min *n*	5-NN OVA wIB4 max *n*	5-NN OVA wIB1w min *n*	5-NN OVA wIB1w max *n*
2 class members	62	67	94	105	88	100	54	66	60	75
3 class members	0	1	3	7	2	7	0	2	0	3
4 class members	0	0	0	0	0	0	0	0	0	0
5 class members	0	0	0	0	0	0	0	0	0	0
6 class members	0	0	0	0	0	0	0	0	0	0
7 non-members	105	111	102	119	97	106	111	124	95	106
total *n* of ties	169	178	205	228	191	211	171	189	158	179

All disease classes were mixed up with Menière's disease: in *ONE1 wscat* from 8.5% (TRA) to 26.0% of the cases (SUD), in *1-NN OVA we* from 4.8% (TRA) to 28.5% (SUD), in *5-NN OVA wscat* from 2.7% (TRA) to 34.0% (ANE) and in *ONE1 experts* from 0% (TRA) to 10.6% (SUD). There were differences in the mixing: *ONE1 wscat*, *1-NN OVA we* and *5-NN OVA wscat* mainly misclassified cases as Menière's diseases, whereas *ONE1 experts* mostly mixed up all classes with benign positional vertigo from 4.3% (SUD) to 30.0% (BRV) and with benign recurrent vertigo from 8.2% (TRA) to 47.1% (VNE). In addition, *ONE1 experts* classified 48.1% of the acoustic neurinoma cases as having sudden deafness. *ONE1 wscat*, *1-NN OVA we* and *5-NN OVA wscat* also mixed up benign recurrent vertigo with benign positional vertigo (27.5%, 44.5% and 44.0% of the cases respectively). *1-NN OVA we* mixed 21.5% of sudden deafness cases with acoustic neurinoma.

Table 4 Confusion matrices of seven disease classes in mean percentages (%) for ONE and the 1- and 5-nearest neighbour OVA methods with the weight sets having the highest total accuracies from 10 times repeated 10-fold cross-validation. Results of *ONE1 experts* added for comparison.

ONE1 wscat: total accuracy 74.6%

	Predicted class						
Correct class	ANE	BPV	MEN	SUD	TRA	VNE	BRV
ANE	**62.3**	0.7	21.2	13.2	0.0	1.5	1.1
BPV	0.0	**55.6**	25.9	1.0	2.7	0.6	14.2
MEN	0.0	0.9	**91.9**	4.0	0.6	0.3	2.3
SUD	2.1	0.0	26.0	**71.9**	0.0	0.0	0.0
TRA	0.0	3.0	8.5	3.6	**83.2**	1.5	0.3
VNE	0.0	5.2	15.7	2.4	1.6	**67.8**	7.2
BRV	0.0	27.5	24.0	0.0	0.0	5.5	**43.0**

1-NN OVA we: total accuracy 74.7%

	Predicted class						
Correct class	ANE	BPV	MEN	SUD	TRA	VNE	BRV
ANE	**67.6**	3.1	25.0	1.6	0.5	2.3	0.0
BPV	0.5	**69.3**	20.4	0.5	1.4	1.3	6.6
MEN	1.5	4.9	**87.2**	0.7	0.9	3.0	1.8
SUD	21.5	0.6	28.5	**45.3**	0.0	4.0	0.0
TRA	1.5	14.7	4.8	0.4	**74.8**	1.4	2.5
VNE	0.0	7.8	11.8	0.1	1.1	**74.4**	4.8
BRV	0.0	44.5	22.5	0.0	0.0	14.0	**19.0**

5-NN OVA wscat: total accuracy 79.7%

	Predicted class						
Correct class	ANE	BPV	MEN	SUD	TRA	VNE	BRV
ANE	**63.1**	1.2	34.0	0.2	0.0	1.5	0.0
BPV	0.5	**70.9**	23.5	0.0	1.7	1.4	2.0
MEN	0.3	2.3	**93.7**	0.3	1.6	1.1	0.7
SUD	2.3	2.1	11.3	**84.3**	0.0	0.0	0.0
TRA	0.0	5.3	2.7	3.7	**86.6**	1.6	0.0
VNE	0.0	8.6	11.0	0.6	1.6	**75.2**	3.1
BRV	0.0	44.0	23.5	0.0	0.0	14.0	**18.5**

ONE1 experts: total accuracy 43.3%

	Predicted class						
Correct class	ANE	BPV	MEN	SUD	TRA	VNE	BRV
ANE	**24.4**	11.5	6.1	48.1	0.0	0.0	9.9
BPV	4.0	**65.9**	5.8	0.6	1.7	1.2	20.8
MEN	7.7	13.4	**42.0**	3.1	1.4	1.4	30.9
SUD	2.1	4.3	10.6	**68.1**	4.3	0.0	10.6
TRA	0.0	24.7	0.0	0.0	**67.1**	0.0	8.2
VNE	1.9	24.8	6.4	1.9	1.9	**15.9**	47.1
BRV	0.0	30.0	5.0	0.0	0.0	0.0	**65.0**

In addition to confounding and overlapping symptoms, patients can actually have two (or more) diseases present simultaneously (Kentala et al., 1996). Furthermore, vertigo diseases resemble each other and can be difficult to differentiate from others, as can be seen in Table 4. Therefore, it is good to check the classification results of ONE with more than one disease suggestion. In the end, the final diagnostic choice must be made by the physician based on the information given on all alternative diseases (Kentala et al., 1996). The classification results when looking for the correct class among the first, second and third diagnosis suggestions given by ONE are given in Table 5. Within the three diagnosis suggestions, the weights computed with the Scatter and IB1w methods improved the total classification accuracy: with the experts' weights the accuracy was 86.2% (*ONE123 experts)* and 90.6% (*ONE123 we*), whereas with the IB1w weights the accuracy was 93.0% (*ONE123 wIB1w*) and with the Scatter weights 94.4% (*ONE123 wscat*). The gap between *ONE123 w1*, *ONE123 experts* and *ONE123 we* narrowed when looking at the three diagnosis suggestions, but *ONE123 w1* was still more robust with a total accuracy of 92.3%. The Scatter-based and IB1w weights also increased the total accuracy compared with the weights 1.

Table 5 The mean true positive rates of seven disease classes and the mean total classification accuracies of the ONE variants having correct diagnosis suggestions within the first, second and third diagnosis suggestions (ONE123) in percentages (%) from 10 times repeated 10-fold cross-validation. The highest TPRs and accuracies are in boldface.

Disease	Cases	ONE123 experts	ONE123 w1	ONE123 we	ONE123 wscat	ONE123 wIB4	ONE123 wIB1w
ANE	131	78.6	90.3	73.9	86.6	82.7	**93.6**
BPV	173	95.4	88.3	85.5	**97.5**	84.6	89.7
MEN	350	78.6	97.9	97.6	**98.1**	95.6	97.7
SUD	47	97.9	98.9	**100.0**	**100.0**	99.6	99.4
TRA	73	**100.0**	**100.0**	94.4	**100.0**	**100.0**	**100.0**
VNE	157	87.9	82.5	**91.7**	85.7	78.0	82.4
BRV	20	**100.0**	77.0	76.0	90.5	99.0	79.0
Median of TPR		95.4	90.3	91.7	**97.5**	95.6	93.6
Total ACC	951	86.2	92.3	90.6	**94.4**	89.5	93.0

Even though the experts could not define weights for vestibulopatia and central lesion, these two classes were used in the classification runs of ONE and the weighted *k*-nearest neighbour method using OVA classifiers. With the machine learning methods we were able to create weights for these two classes and were thus able to use nine disease classes in the classification runs. When comparing the classification results of nine disease classes with ONE and the attribute weighted 1- and 5-nearest neighbour methods (Table 6), the best results were achieved with the 5-nearest neighbour method with the weights calculated by the Scatter method (*5-NN OVA wscat*). It classified 73.3% of cases correctly, whereas other *wk*-NN OVA methods recognized 62.9% to 70.1% and ONE variants 59.1% to 62.4% cases correctly. *5-NN OVA wscat* also had the highest Cohen's kappa value (0.66). The highest median TPR (65.7%) was yielded with *ONE1 wIB4*.

Table 6 The mean true positive rates of nine disease classes and the mean total classification accuracies of ONE's first diagnosis suggestions (ONE1) and the attribute weighted *k*-nearest neighbour method with OVA (*wk*-NN OVA) in percentages (%) from 10 times repeated 10-fold cross-validation. In addition, Cohen's Kappa (*K*) and the Kappa Chance agreement (P_c) are presented. The highest TPRs and accuracy are in boldface.

						wk-NN OVA w1		*wk*-NN OVA wscat		*wk*-NN OVA wIB4		*wk*-NN OVA wIB1w	
Disease	Cases	ONE1 w1	ONE1 wscat	ONE1 wIB4	ONE1 wIB1w	1-NN	5-NN	1-NN	5-NN	1-NN	5-NN	1-NN	5-NN
ANE	131	65.6	62.7	66.2	**66.6**	64.7	61.6	60.6	60.0	60.3	57.1	59.0	56.4
BPV	173	32.6	31.4	25.1	29.4	57.7	64.6	57.5	**65.1**	60.2	64.5	58.6	60.0
MEN	350	81.3	80.2	65.7	79.3	78.4	**94.4**	77.3	93.1	81.3	93.5	86.0	94.0
SUD	47	61.3	68.9	76.2	63.0	37.0	28.3	61.5	**81.5**	20.2	23.0	24.0	28.7
TRA	73	69.6	77.3	**95.6**	79.3	59.2	73.4	73.8	85.2	61.4	73.3	47.4	55.5
VNE	157	63.9	64.9	58.3	64.3	62.7	72.7	65.8	**75.4**	64.8	73.2	64.6	70.4
BRV	20	4.0	4.0	**20.5**	3.0	19.5	12.5	19.0	14.5	11.0	14.0	13.5	17.0
VES	55	40.2	41.3	**47.5**	42.4	36.4	25.3	36.9	26.7	32.5	27.6	30.0	17.5
CL	24	46.7	46.7	**89.2**	40.8	23.3	7.5	22.9	7.5	16.7	7.5	12.5	7.5
Median of TPR		61.3	62.7	**65.7**	63.0	57.7	61.6	60.6	65.1	60.2	57.1	47.4	55.5
Total ACC	1030	62.2	62.4	59.1	61.9	62.9	70.1	64.6	**73.3**	62.9	69.2	63.0	66.6
K		0.54	0.54	0.52	0.54	0.54	0.61	0.56	**0.66**	0.53	0.60	0.53	0.56
P_c		0.18	0.18	**0.15**	0.18	0.20	0.23	0.19	0.22	0.21	0.23	0.22	0.24

The highest total accuracy of the ONE variants at 62.4% was achieved with *ONE1 wscat,* having a kappa value of 0.54 and a median TPR of 62.7%. Other machine learnt weights (IB4 and IB1w) slightly reduced the total accuracy compared with the equally weighted ONE and 5-NN OVA. The weights based on the Scatter method seemed to work with all methods: ONE and 1- and 5-NN OVA with the Scatter-based weights had the highest total accuracies within the methods.

The two added classes (vestibulopatia and central lesion) were difficult to recognize with both the attribute weighted *k*-nearest neighbour OVA methods and ONE (Table 6). Vestibulopatia was correctly classified with the weighted *k*-NN OVA combinations from 17.5% to 36.9% of the cases and with the first suggestion of ONE's weight combinations from 40.2% to 47.5% of the cases. The classification of central lesion was not much easier for the weighted *k*-NN OVA: from 7.5% to 23.3% of the cases were correctly classified with the weighted *k*-NN OVA combinations. Instead, ONE classified from 40.8% to 89.2% of the central lesion cases correctly. Furthermore, the addition of these two difficult diseases to the classification reduced the true positive rates of the other seven classes with some methods, especially with benign recurrent vertigo (39.0% decrease with *ONE1 wscat*), benign positional vertigo (25.1% decrease with *ONE1 wIB4*), and Menière's disease (15.7% decrease with *ONE1 wIB4*) (Tables 2 and 6).

With the nine disease classes, the total number of cases having tied voting 1- and 5-nearest neighbour method OVA classifiers within the 10 times repeated 10-fold cross-validations (Table 7) increased compared with the seven disease classes. However, ONE did not have more than one case having the same highest score and the same max-min score difference for two class suggestions. The total number of ties occurring within the cross-validation runs varied with the 1-nearest neighbour OVA method from 270 to 332 (26.2% to 32.2%) and with the 5-nearest neighbour OVA method from 244 to 290 (23.7% to 28.2%). The weighted 1- and 5-nearest neighbour OVA method having the lowest total number of tied voting classifiers was achieved with *1-NN OVA wIB4* (270 to 302 ties) and *5-NN OVA wIB1w* (244 to 267 ties). Interestingly, the proportion of non-member voting classifiers with *1-NN OVA* stayed almost the same with nine disease classes, whereas the proportion increased with *5-NN OVA*: during the classification of nine diseases with *1-NN OVA* there were at worst 14.1% non-member voting classifiers (14.5% with seven diseases) and 20.8% with *5-NN OVA* (13.0% with seven diseases).

Table 7 The minimum and maximum number (*n*) of cases having tied diagnosis suggestions or a tied voting situation occurring in 10 times repeated 10-fold cross-validation with nine disease classes (the number of cases covers the entire 10-fold data).

(a) diagnosis suggestions of ONE with the same highest score and maximum score and minimum score difference.

n of tied suggestions	ONE1 w1 min *n*	ONE1 w1 max *n*	ONE1 wscat	ONE1 wIB4	ONE1 wIB1w
2 suggestions	0	1	0	0	0
total *n* of ties	0	1	0	0	0

(b) tied voting with the attribute weighted 1- and 5-nearest neighbour methods with OVA classifiers.

n of tied voting classifiers	1-NN OVA w1		1-NN OVA wscat		1-NN OVA wIB4		1-NN OVA wIB1w	
	min *n*	max *n*	min *n*	max *n*	min *n*	max *n*	min *n*	max *n*
2 class members	152	169	161	180	125	150	154	168
3 class members	6	11	9	15	7	13	6	11
4 class members	0	2	0	2	0	0	0	1
5 class members	0	0	0	0	0	0	0	0
6 class members	0	0	0	0	0	0	0	0
7 class members	0	0	0	0	0	0	0	0
8 class members	0	0	0	0	0	0	0	0
9 non-members	117	134	124	140	125	145	110	126
total *n* of ties	292	306	311	332	270	302	280	300

n of tied voting classifiers	5-NN OVA w1		5-NN OVA wscat		5-NN OVA wIB4		5-NN OVA wIB1w	
	min *n*	max *n*	min *n*	max *n*	min *n*	max *n*	min *n*	max *n*
2 class members	54	67	87	100	53	65	58	74
3 class members	0	1	3	6	0	1	0	2
4 class members	0	0	0	0	0	0	0	0
5 class members	0	0	0	0	0	0	0	0
6 class members	0	0	0	0	0	0	0	0
7 class members	0	0	0	0	0	0	0	0
8 class members	0	0	0	0	0	0	0	0
9 non-members	187	209	171	185	197	214	175	199
total *n* of ties	253	272	269	290	256	273	244	267

The mean confusion matrices for the classification methods ONE and 1- and 5-nearest neighbour methods using OVA classifiers with weight combinations that had the highest total accuracy from the 10 times repeated 10-fold cross-validation within nine disease classes are given in Table 8. All disease classes were again mixed up with Menière's disease. In particular, the cases of sudden deafness were classified as Menière's disease: in *ONE1 wscat* 25.7%, in *1-NN OVA wscat* 34.7% and in *5-NN OVA wscat* 53.6%. The 1- and 5-nearest neighbour methods using the Scatter weights mixed up the cases of vestibulopatia, central lesion and benign recurrent vertigo with benign positional vertigo besides Menière's disease. In addition, benign recurrent vertigo was badly mixed up with vestibulopatia with all three methods: in *ONE1 wscat* 62.0%, in *1-NN OVA wscat* 28.0% and in *5-NN OVA wscat* 32.5%. With *ONE1 wscat,* the cases of benign positional vertigo were mixed up with vestibulopatia, central lesion and Menière's disease.

When looking the correct class within the three best diagnosis suggestions of ONE with the nine disease classes (Table 9), the best total accuracy was achieved with *ONE123 wscat* (85.0%). *ONE123 w1* was the second best with 84.9% total accuracy and *ONE123 wIB1w* was the third best with 84.7% accuracy. However, the highest median TPR (91.2%) was achieved with *ONE123 wIB4*. The addition of two disease classes

reduced the true positive rates of the other seven classes (Tables 5 and 9). The TPRs reduced at worst by 31.2% (BPV with *ONE123 wscat*) and 30.0% (BRV with *ONE123 w1*) within the three first diagnosis suggestions compared with results of ONE when using seven disease classes in the knowledge base.

Table 8 Confusion matrices of nine disease classes in mean percentages (%) for the ONE and the 1- and 5-nearest neighbour OVA methods with the weight sets having the highest total accuracies within the methods from 10 times repeated 10-fold cross-validation.

ONE1 wscat: total accuracy 62.4%

	Predicted class								
Correct class	ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL
ANE	**62.7**	0.0	19.7	12.3	0.0	1.3	0.2	2.9	0.9
BPV	0.0	**31.4**	14.2	0.9	1.6	0.6	3.9	32.9	14.6
MEN	0.0	0.3	**80.2**	3.9	0.6	0.3	0.9	4.7	9.0
SUD	2.1	0.0	25.7	**68.9**	0.0	0.0	0.0	2.1	1.1
TRA	0.0	1.2	6.3	3.4	**77.3**	1.4	0.0	5.8	4.7
VNE	0.0	1.3	10.3	2.3	0.7	**64.9**	1.8	12.9	5.8
BRV	0.0	9.0	12.0	0.0	0.0	5.0	**4.0**	62.0	8.0
VES	0.0	4.9	21.8	0.0	0.0	0.0	8.9	**41.3**	23.1
CL	0.0	0.0	16.7	0.0	4.2	4.2	0.0	28.3	**46.7**

1-NN OVA wscat: total accuracy 64.6%

	Predicted class								
Correct class	ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL
ANE	**71.5**	1.7	22.1	2.7	0.0	1.6	0.0	0.1	0.3
BPV	0.5	**65.1**	17.5	0.5	1.6	1.5	2.5	8.8	2.0
MEN	1.1	4.8	**86.4**	1.2	0.6	0.9	1.5	2.7	0.7
SUD	10.0	1.5	34.7	**48.3**	0.0	3.6	0.0	1.9	0.0
TRA	0.0	4.1	4.1	0.0	**87.7**	4.1	0.0	0.0	0.0
VNE	0.0	3.2	8.6	0.6	0.9	**79.4**	1.7	5.1	0.6
BRV	0.0	25.5	18.0	0.0	0.0	9.0	**14.5**	28.0	5.0
VES	1.8	21.5	22.9	0.0	0.0	3.6	11.6	**28.9**	9.6
CL	0.0	31.2	26.3	0.0	5.0	5.0	4.2	19.2	**9.2**

5-NN OVA wscat: total accuracy 73.3%

	Predicted class								
Correct class	ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL
ANE	**68.7**	0.6	26.6	2.0	0.0	1.8	0.0	0.2	0.2
BPV	0.5	**64.6**	25.1	0.0	1.0	0.7	0.6	6.6	0.9
MEN	0.0	2.1	**95.3**	0.0	0.5	0.8	0.3	1.0	0.0
SUD	13.2	0.9	53.6	**28.7**	0.0	2.8	0.0	0.2	0.6
TRA	0.0	4.7	7.1	0.0	**83.2**	0.5	2.3	2.2	0.0
VNE	0.0	5.2	12.7	0.0	0.7	**77.6**	0.4	3.2	0.0
BRV	0.0	33.5	19.0	0.0	0.0	5.0	**10.0**	32.5	0.0
VES	1.3	28.5	34.5	0.5	0.0	0.4	3.1	**27.3**	4.4
CL	0.0	21.7	44.2	0.0	4.2	4.2	0.4	22.5	**2.9**

Table 9 The mean true positive rates of nine disease classes and the mean total classification accuracies of ONE variants having correct diagnosis suggestion within the first, second and third diagnosis suggestions (ONE123) in percentages (%) from 10 times repeated 10-fold cross-validation. The highest TPRs and accuracies are in boldface.

Disease	Cases	ONE123 w1	ONE123 wscat	ONE123 wIB4	ONE123 wIB1w
ANE	131	**87.3**	80.5	75.7	86.1
BPV	173	63.9	**66.3**	61.0	64.3
MEN	350	**97.4**	96.7	91.2	96.9
SUD	47	92.6	**97.0**	96.8	92.1
TRA	73	96.7	96.3	**99.7**	98.2
VNE	157	**73.6**	72.4	70.0	72.6
BRV	20	47.0	68.5	**74.0**	59.0
VES	55	90.9	**95.8**	93.6	89.5
CL	24	80.8	86.7	**97.9**	80.8
Median of TPR		87.3	86.7	**91.2**	86.1
Total ACC	1030	84.9	**85.0**	81.7	84.7

5 Conclusions

The Scatter method and the weight calculation method of the instance-based learning method with two variants (IB4 and IB1w) were used in the attribute weight calculation. The created attribute weights were tested with the nearest pattern method of ONE and the attribute weighted *k*-nearest neighbour method with One-vs-All classifiers. The expert-defined weights and weights set to 1 were also used in the classification.

Previous study (Varpa et al., 2008) showed that learning fitness values for attribute values with the machine learning method improved the classification of ONE. However, there was a need for attribute weighting in order to ameliorate the discrimination of the classes: some classes were mixed up with other classes when having equal attribute weighting (all weights set to 1). Nevertheless, as the results of this study show, attribute weighting is a demanding task and does not always help recognition. The Scatter-based weights were the only machine learnt weights that improved the total accuracies compared with the equal weighting. The IB4 and IB1w weights did not help the separation of classes with the attribute weighted *k*-nearest neighbour OVA method and ONE. Overall, the best total accuracy was achieved with the attribute weighted 5-nearest neighbour OVA method using the Scatter weights.

Based on the total accuracies and the Cohen's kappa values, the machine learnt weights improved the classification of ONE compared with the weights defined by the experts when classifying seven disease classes. The Scatter-based weights yielded the best total accuracy and Cohen's kappa for ONE (74.6% and 0.67). ONE with the weights set to 1 classified cases better than ONE with the experts' weights. With the attribute weighted 1-nearest neighbour OVA method, the best total accuracy and Cohen's kappa were achieved with the experts' weights (74.7% and 0.67), whereas with the attribute weighted 5-nearest neighbour OVA method, the best total accuracy and Cohen's kappa were yielded with the Scatter-based weights (79.7% and 0.73). Also, with nine disease classes, the best total accuracy and Cohen's kappa with ONE (62.4% and 0.54) and with attribute weighted 1- and 5-nearest neighbour OVA methods (64.6% and 0.56 and 73.3%

and 0.66 respectively) were achieved using the Scatter-based weights. Thus, the weights based on the Scatter method worked well with both weight utilizing methods. The highest true positive rates within the disease classes varied depending on the utilised inference mechanism and class: in some disease classes even the weights set to 1 or the weights defined by the experts produced the best accuracy.

When adding two difficult diseases (vestibulopatia and central lesion) to the knowledge base of ONE, the true positive rates of the other seven disease classes decreased considerably, especially with the diseases benign recurrent vertigo, benign positional vertigo and Menière's disease. The decrease can also be seen in the results of the attribute weighted *k*-nearest neighbour with OVA classifiers. This confirms that certain disease classes have overlapping and confounding symptoms (Kentala et al., 1998), and, therefore, are mixed up with other diseases during classification.

The kappa chance value P_c describes the "agreement" probability that can really be attributed to chance alone (Ben-David, 2007). In Ben-David's research, the average kappa chance within different classification methods (C4.5, sequential minimal optimisation , Naïve Bayes, logistic regression and random forest) tested with different data sets from the UCI Machine Learning Repository were 0.35, thus showing that more than one-third of the hits in the classification results could not be attributed to the classifiers' sophistication. Compared with this average kappa chance value, ONE and the attribute weighted *k*-nearest neighbour OVA methods do not seem to let chance affect the classification results as much. The kappa chance values varied with ONE from 0.15 to 0.24 with seven diseases, from 0.15 to 0.18 with nine diseases and with the weighted 1- and 5-nearest neighbour methods from 0.23 to 0.26 and 0.19 to 0.24 respectively.

Otoneurology is a difficult domain: there are many reasons for vertigo and some diseases are considered challenging to diagnose because of the overlapping and similar symptoms within diseases. Therefore, physicians see tools that support making a diagnosis as very useful (Aalto, 2005). In order to support more diagnosing, we are aiming to make ONE a hybrid decision support system - *i.e.*, to use several inference methods while making diagnosis suggestions. With more than one inference method it is possible to make more reliable decisions. Therefore, in this research we used the attribute weighted *k*-nearest neighbour OVA method with ONE's classification method. ONE and the attribute weighted *k*-nearest neighbour OVA method have different approaches to the classification problem: ONE handles descriptions of the diseases and can advise the user why the diseases could be possible or not (*e.g.*, do the occurring symptoms fit the disease and what tests need to be done in order to confirm the diagnosis), whereas the attribute weighted *k*- nearest neighbour OVA method handles cases individually and classifies new cases based on their *k* most similar neighbours giving information about similar cases.

The next step in the attribute weighting of ONE is to use more adaptive machine learning methods in the attribute weight calculation. In our next study, we will use a genetic algorithm (Michalewicz, 1992; Mitchell, 1996) as an adaptive weight calculation method. This approach has been shown to improve the results with a *k*-nearest neighbour classifier (Kelly and Davis, 1991).

As the results showed, it is important to have appropriate attribute weights. The extent of the effect the attribute weights had on the classification results depended on the classification method used. Based on the Cohen's kappa evaluations, the ONE method is more sensitive to the attribute weights. The attribute weighted 5-nearest neighbour OVA variants with different weight sets agreed more with each other than attribute weighted 1-nearest neighbour OVA and ONE with different weight sets.

The machine learning methods for weight calculation described in this study are not domain-dependent and can be applied in totally different domains. The only

prerequisite is that there is enough data in order to apply machine learning methods in attribute weight calculation. In the future, the attribute weighting methods will be tested with several data sets from different domains. Also other attribute weighting and weighted classification methods will be taken into use in further research.

Acknowledgements

Kirsi Varpa acknowledges the support of The Tampere Doctoral Programme in Information Science and Engineering (TISE), Tampere University, The Ella and Georg Ehrnrooth Foundation, Finnish Cultural Foundation, Päijät-Häme Regional fund, The Onni and Hilja Tuovinen Foundation and Oskar Öflund's Foundation , who granted scholarships for postgraduate studies.

The authors are grateful to Docent E. Kentala, M.D., and Prof. I. Pyykkö, M.D., for their help in collecting the otoneurological data and medical advice. The authors acknowledge also the IT Center for Science (CSC) whose supercomputer resources were used in some cross-validation runs of IB4 and IB1w.

References

Aalto, P. (2005) *Equihear-markkinaselvitys – Kuulon ja huimauksen IT-pohjainen konsepti* (in Finnish), Finn-Medi Tutkimus report, Tampere, Finland.

Aha, D.W. (1992) 'Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms', *International Journal of Man-Machine Studies*, Vol. 36, No. 2, pp.267–287.

Aha, D.W., Kibler, D. and Albert, M.K. (1991) 'Instance-based learning algorithms', *Machine Learning*, Vol. 6, No. 1, pp.37–66.

Auramo, Y. and Juhola, M. (1995) 'Comparison of inference results of two otoneurological expert systems', *International Journal of Bio-Medical Computing*, Vol. 39, No. 3, pp.327–335.

Auramo, Y. and Juhola, M. (1996) 'Modifying an expert system construction to pattern recognition solution', *Artificial Intelligence in Medicine*, Vol. 8, pp.15–21.

Auramo, Y., Juhola, M. and Pyykkö, I. (1993) 'An expert system for the computer-aided diagnosis of dizziness and vertigo', *Medical Informatics*, Vol. 18, No. 4, pp.293–305.

Ben-David, A. (2007) 'A lot of randomness is hiding in accuracy', *Engineering Applications of Artificial Intelligence*, Vol. 20, No. 7, pp.875–885.

Blum, A.L. and Langley, P. (1997) 'Selection of relevant features and examples in machine learning', *Artificial Intelligence,* Vol. 97, No. 1–2, pp.245–271.

Cardie, C. and Howe, N. (1997) 'Improving minority class prediction using case-specific feature weights', in Fisher, D.H. (Ed.), *ICML 1997: Proceedings of the 14th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, pp.57–65.

Cohen, J. (1960) 'A coefficient of agreement for nominal scales' *Educational and Psychological Measurement*, Vol. 20, No. 1, pp.37–46.

Cover, T.M. and Hart, P.E. (1967) 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory,* Vol. 13, No. 1, pp.21–27.

Debnath, S., Ganguly, N. and Mitra, P. (2008) 'Feature weighting in content based recommendation system using social network analysis', in *WWW2008: Proceedings of the 17th International Conference on World Wide Web*, ACM, New York, pp.1041–1042.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H. and Herrera, F. (2011) 'An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes', *Pattern Recognition*, Vol. 44, No. 8, pp.1761–1776.

Hall, M. (2007) 'A decision tree-based attribute weighting filter for naïve Bayes', *Knowledge-Based Systems*, Vol. 20, No. 2, pp.120–126.

Havia, M. (2004) '*Menière's Disease Prevalence and Clinical Picture'*. Academic dissertation, Department of Otorhinolaryngology, University of Helsinki, Finland. http://ethesis.helsinki.fi/julkaisut/laa/kliin/vk/havia/menieres.pdf (Accessed 5th March 2015).

Juhola, M. and Siermala, M. (2012) 'A scatter method for data and variable importance evaluation', *Integrated Computer-Aided Engineering*, Vol. 19, pp.137–149.

Kelly, J.D. and Davis, L. (1991) 'A hybrid genetic algorithm for classification', in: *IJCAI 1991: Proceedings of the 12th International Joint Conference on Artificial Intelligence vol. 2,* San Franciso, CA, USA, Morgan Kaufmann, pp.645–650.

Kentala, E. (1996) 'Characteristics of six otologic diseases involving vertigo', *American Journal of Otolaryngology*, Vol. 17, No. 6, pp.883–892.

Kentala, E., Auramo, Y., Juhola, M. and Pyykkö, I. (1998) 'Comparison between diagnoses of human experts and a neurotologic expert system,' *Annals of Otology, Rhinology and Laryngology*, Vol. 107, No. 2, pp.135–140.

Kentala, E., Pyykkö, I., Auramo, Y. and Juhola, M. (1995) 'Database for vertigo', *Otolaryngology – Head and Neck Surgery*, Vol. 112, No. 3, pp.383–390.

Kentala, E., Pyykkö, I., Auramo, Y. and Juhola, M. (1996) 'Otoneurological expert system', *Annals of Otology, Rhinology and Laryngology*, Vol. 105, No. 8, pp.654–658.

Kira, K. and Rendell, L.A. (1992) 'A practical approach to feature selection', in *ICML 1992: Proceedings of the 9th International Conference on Machine Learning*, Morgan Kaufmann, Scotland, pp.249–256.

Kohavi, R. and John, G. (1997) 'Wrappers for feature subset selection', *Artificial Intelligence*, Vol. 97, No. 1–2, pp.273–324.

Landis, J.R. and Koch, G.G. (1977) 'The measurement of observer agreement for categorical data', *Biometrics*, Vol. 33, No. 1, pp.159–174.

Laurikkala, J., Kentala, E., Juhola, M., Pyykkö, I. and Lammi, S. (2000) 'Usefulness of imputation for the analysis of incomplete otoneurological data', *International Journal of Medical Informatics*, Vol. 58–59, pp.235–242.

Lee, H., Kim, E. and Park, M. (2007) 'A genetic feature weighting scheme for pattern recognition', *Integrated Computer-Aided Engineering,* Vol. 14, No. 2, pp.161–171.

Michalewicz, Z. (1992) *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, Berlin Heidelberg, New York, 1992.

Mitchell, M. (1996) *An Introduction to Genetic Algorithms*, MIT Press, Cambridge.

Mitchell, T. (1997) *Machine Learning*, McGraw-Hill, New York.

Rifkin, R. and Klautau, A. (2004) 'In defense of one-vs-all classification', *Journal of Machine Learning Research*, Vol. 5, pp.101–141.

Saeys, Y., Inza, I. and Larrañaga, P. (2007) 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, Vol. 23, No. 19, pp.2507–2517.

Schulerud, H. and Albregtsen, F. (2004) 'Many are called, but few are chosen. Feature selection and error estimation in high dimensional spaces', *Computer Methods and Programs in Biomedicine*, Vol. 73, No. 2, pp.91–99.

Siermala, M., Juhola, M., Laurikkala, J., Iltanen, K., Kentala, E. and Pyykkö, I. (2007) 'Evaluation and classification of otoneurological data with new data analysis methods based on machine learning' *Information Sciences*, Vol. 177, No. 9, pp.1963–1976.

Varpa, K., Iltanen, K. and Juhola, M. (2008) 'Machine learning method for knowledge discovery experimented with otoneurological data', *Computer Methods and Programs in Biomedicine*, Vol. 91, No. 2, pp.154–164.

Varpa, K., Iltanen, K., Juhola, M., Kentala, E. and Pyykkö, I. (2006) 'Refinement of the otoneurological decision support system and its knowledge acquisition process', in Engelbrecht, R. and Hasman, A. (Eds.): *MIE2006: Proceedings of the 20th International Congress of the European Federation for Medical Informatics*. Maastricht, pp.97–202.

Viikki, K., '*Machine Learning on Otoneurological Data: Decision Trees for Vertigo Diseases',* Academic dissertation, Department of Computer Sciences, University of Tampere, Finland, 2002. http://urn.fi/urn:isbn:951-44-5390-5 (Accessed 5th March 2015).

Vivencio, D.P., Hruschka Jr., E.R., do Carmo Nicoletti, M., dos Santos, E.B. and Galvão, S.D.C.O. (2007) 'Feature-weighted k-nearest neighbor classifier', in *FOCI 2007: Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence*, pp.481–486.

Wettschereck, D., Aha, D.W., and Mohri, T. (1997) 'A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms', *Artificial Intelligence Review*, Vol. 11, no. 1–5, pp.273–314.

Wettschereck, D. and Aha, D.W. (1995) 'Weighting features', in Veloso, M. and Aamodt, A. (Eds.): *ICCBR 1995: Proceedings of the 1st International Conference on Case-Based Reasoning Research and Development,* Springer-Verlag, London, pp.347–358.

Wilson, R.D. and Martinez, T.R. (1997) 'Improved heterogeneous distance functions', *Journal of Artificial Intelligence Research*, Vol. 6, pp.1–34.

Zeng, X. and Martinez, T.R. (2004) 'Feature weighting using neural networks', in *Proceedings of 2004 IEEE International Joint Conference on Neural Networks, Vol. 2*, pp.1327–1330.

PUBLICATION IV

Genetic Algorithm Based Approach in Attribute Weighting for a Medical Data Set

Kirsi Varpa, Kati Iltanen and Martti Juhola

Journal of Computational Medicine 2014, 2014, pp. 1–11
https://doi.org/10.1155/2014/526801

Hindawi Publishing Corporation
Journal of Computational Medicine
Volume 2014, Article ID 526801, 11 pages
http://dx.doi.org/10.1155/2014/526801

Research Article

Genetic Algorithm Based Approach in Attribute Weighting for a Medical Data Set

Kirsi Varpa, Kati Iltanen, and Martti Juhola

Computer Science, School of Information Sciences, University of Tampere, 33014 Tampere, Finland

Correspondence should be addressed to Martti Juhola; martti.juhola@sis.uta.fi

Received 28 May 2014; Revised 30 July 2014; Accepted 6 August 2014; Published 3 September 2014

Academic Editor: Martin J. Murphy

Genetic algorithms have been utilized in many complex optimization and simulation tasks because of their powerful search method. In this research we studied whether the classification performance of the attribute weighted methods based on the nearest neighbour search can be improved when using the genetic algorithm in the evolution of attribute weighting. The attribute weights in the starting population were based on the weights set by the application area experts and machine learning methods instead of random weight setting. The genetic algorithm improved the total classification accuracy and the median true positive rate of the attribute weighted k-nearest neighbour method using neighbour's class-based attribute weighting. With other methods, the changes after genetic algorithm were moderate.

1. Introduction

One of the most commonly used simple classification methods is the nearest neighbour (NN) method that classifies a new case into the class of its nearest neighbour case [1]. The nearest neighbour method is an instance-based learning method that searches for the most similar case of the test case from the training data by some distance measure, usually with the Euclidean distance. A natural extension to NN is the k-nearest neighbour (k-NN) method that assigns the majority class of the k nearest training cases for the test case [2]. Different refinements and extensions have been proposed for k-NN in order to improve classification results and overcome classification problems, for example, distance-weighting of neighbours [2], extensions using properties of the data set [3], weighting of attributes [2, 4, 5], and attribute weight optimization with genetic algorithms (GA) [6–11].

Genetic algorithms [12, 13] and other evolution algorithms [14, 15] have been utilized in various complex optimization and simulation problems because of their powerful search and optimization capabilities. A search method of a genetic algorithm is a combination of directed and stochastic search and the search can be done multidirectionally because GA maintains a population of potential solutions from the search space [14]. The basics of the search method of GA underlie in natural selection and genetic inheritance [12]; individuals of the population are used in the reproduction of new solutions by means of crossover and mutation. Genetic algorithms have been used with various machine learning methods to optimize weighting properties of the method. Since our research is based on the nearest neighbour search applying machine learning methods, we concentrate on related works where GAs have been applied only with the k-nearest neighbour method. Kelly and Davis [6] combined the GA with a weighted k-nearest neighbour (wk-NN) method in the algorithm called GA-WKNN in order to find a single attribute weight vector that would improve the classification results of the wk-NN. A similar kind of approach was used in [7] where GA was combined with the wk-NN and a parallel processing environment in order to optimize classification of large data sets. In both studies, a set of real-valued weights for attributes to discriminate all classes of data were achieved as a result after GA runs. The study of Hussein et al. [8] showed that GA can be applied successfully in setting a real-valued weight set for 1-NN classifier but the improvement of accuracy happened at the expense of increase in processing time. Results showed that GA methods combining the wk-NN outperformed the basic k-NN [6–8]. However, a single set of weights for all classes is not always the best solution because attributes have a different effect on classes [11]. Therefore,

solutions for searching for a weight for each class and attribute have been developed. Lee et al. [9] combined the GA-based attribute weighting method with a modified k-NN, thus, forming an adaptive feature weighting method A3FW-MNN that used different sets of attribute weights for different classes. Also, Mateos-García et al. [10] assigned different weights to every attribute depending on each class in their evolutionary algorithm called Label Dependent Feature Weighting (LDFW) algorithm.

In this research we studied whether the classification performance of the attribute weighted machine learning methods based on the nearest neighbour search can be improved when using the genetic algorithm in the evolution of attribute weighting based on the experts and machine learning methods when runs were made with a medical data set. This medical data has been our test data in our previous researches [16, 17].

2. Material

In this research an otoneurological data set having 951 cases from seven different vertigo diseases (classes) (Table 1) was used. The data was collected over a decade starting from the 1990s in the Department of Otorhinolaryngology at Helsinki University Central Hospital, Finland, where experienced specialists confirmed all the diagnoses. The distribution of the disease classes is imbalanced; over one-third of the cases belong to the Menière's disease class (36.8%), whereas the smallest disease class benign recurrent vertigo has only 2.1% of the cases.

In total, the data includes 176 attributes concerning a patient's health status: occurring symptoms, medical history, and clinical findings in otoneurologic, audiologic, and imaging tests [18, 19]. Clinical testing has not been done to every patient and, therefore, there are several test results that have missing values of the attributes. Attributes with low frequencies of available values were left outside this research. After leaving out the attributes having over 35% missing values, 94 attributes remained to be used in this research: 17 quantitative (integer or real value) and 77 qualitative attributes (of which 54 were binary (yes/no), 20 were ordinal, and 3 were nominal). Genetic algorithm runs were done with the data including missing attribute values.

Table 1: The frequency distribution of vertigo disease classes.

	Disease name	Abbreviation	Frequency	%
1	Acoustic neurinoma	ANE	131	13.8
2	Benign positional vertigo	BPV	173	18.2
3	Menière's disease	MEN	350	36.8
4	Sudden deafness	SUD	47	4.9
5	Traumatic vertigo	TRA	73	7.7
6	Vestibular neuritis	VNE	157	16.5
7	Benign recurrent vertigo	BRV	20	2.1
	Total		951	100

3. Genetic Algorithm

The basic idea of the genetic algorithm is the following: in the beginning, a population of individuals is formed either randomly or with information about the application domain. Traditionally, a binary representation of the individuals has been used but in multidimensional and numerical problems real-valued representation is nowadays used [14]. In each generation, the individuals of the population are evaluated with an objective evaluation function, thus, giving the individual its fitness rate. A selection method is used to find the fittest individuals for a new population. Some individuals of the new population undergo reproduction by means of crossover and mutation. In the crossover, the information of the individuals is swapped in their corresponding elements. Mutation alters one or more elements of the individual arbitrarily. Elitism is a commonly applied survivor selection method. It keeps the current fittest individual unchanged in the population so the high-performance individuals are not lost from one generation to the next [20]. The GA can be ended after a fixed number of iterations or if no further improvement is observed after some number of generations.

We utilized the genetic algorithm in the evolution of the attribute weight values. A pseudocode of the used genetic algorithm is given in Pseudocode 1. A population contained 21 individuals that used real-valued representation instead of binary presentation because the attribute weight values were described with real-valued numbers, not just with 0 and 1. Each individual consisted of seven different attribute weight sets for 94 attributes. The individuals of the starting population were based on the weights set by the experts and machine learning methods. The starting population is defined more accurately in Section 3.1. The genetic algorithm used a roulette-wheel selection in parent selection and a uniform crossover with discrete recombination in offspring creation. The crossover was done in 80.0% probability ($p_c = 0.8$) and the crossover points were selected randomly and independently for each gene (a field on an individual). Mutation was done in 1.0% probability ($p_m = 0.01$) for the gene and it was done also in a uniform manner: a random value was drawn from the range [0, 1] which was set as a new value in the current position. In addition, elitism was used in order to keep the best individual within the population during runs. We did not want to lose the best performing weight set during the evolution. If the number of the individuals was higher than 21 in the end of the generation, a survivor selection was used. The individuals were ordered by their classification performance and the individuals with the lowest accuracy were discarded from the population. The genetic algorithm ended after 20 generations or if the best classification accuracy maintained the same during 10 successive generations. Furthermore, if all the individuals were the same in the population, the evaluation ended. The parameters used in the GA runs are described in Table 2.

The genetic algorithm runs were done separately with three different machine learning methods used in the population evaluation: with the nearest pattern method of the otoneurological expert system (ONE), with the

$$\text{data } D = \begin{bmatrix} Case_1 \\ \vdots \\ Case_{951} \end{bmatrix} = \begin{bmatrix} [c_{1,1}, \ldots, c_{1,94}] \\ \vdots \\ [c_{951,1}, \ldots, c_{951,94}] \end{bmatrix}$$

$$\text{population } Weights = \begin{bmatrix} Weight_1 \\ \vdots \\ Weight_{21} \end{bmatrix} = \begin{bmatrix} [w_{1,1,1}, \ldots, w_{1,1,94}; \ldots; w_{1,7,1}, \ldots, w_{1,7,94}] \\ \vdots \\ [w_{21,1,1}, \ldots, w_{21,1,94}; \ldots; w_{21,7,1}, \ldots, w_{21,7,94}] \end{bmatrix}$$

$population\_size = 21$
$p_c = 0.8$ //Crossover rate
$p_m = 0.01$ //Mutation rate
divide data D into 10 equally-sized subsets
for $cv\_round = 1$ to 10 **do**
    divide *training* data D-d_{cv_round} into *train* (6 subsets) and *test* (3 subsets) data
    initialize methods with *train* data:
        cwk-NN and *wk*-NN OVA: HVDM initialization
        ONE: fitness value calculation for values of attributes
    evaluate starting population *Weights* with *test* data and ONE/*cwk*-NN/*wk*-NN OVA
    while ending terms of GA are not fulfilled **do**
        //Survivor selection: Elitism
        search for the individual with the highest fitness rate from the population
        //Parent selection: Roulette-wheel selection with fitness-proportionate selection
        for each individual in the population **do**
            calculate individual's fitness proportionate rate = individual's fitness rate/sum of individuals' fitness rates
            calculate individual's cumulative fitness proportionate rate
        end for
        while nr of individuals in the mating pool is smaller than *population_size* **do**
            generate a random number r from [0, 1]
            search for the jth individual that has smaller cumulative fitness proportionate rate than r
            add the jth individual in the mating pool
        end while
        //Crossover: Uniform crossover with discrete recombination
        for each individual in the mating pool **do**
            generate a random number s from [0, 1]
            if s is smaller than p_c **then**
                add the individual in the parent pool
            else
                add the individual in the new population (offspring is a direct copy of its parent)
            end if
        end for
        while two individuals can be taken from the parent pool **do**
            if two individuals are exactly the same **then**
                add the first individual into the new population
                take new individual from the parent pool to use in the crossover
            end if
            for each disease class weight set **do**
                select the crossover points randomly
                swap information of two individuals in the corresponding crossover points (create children)
            end for
            add children in the new population
        end while
        //Mutation: Uniform mutation
        for each individual in the new population **do**
            for each gene of individual **do**
                generate a random number t from [0, 1]
                if t is smaller than p_m **then**
                    select a random value v from the range [0, 1]
                    set the value v as a new value of the gene
                end if
            end for
        end for

Pseudocode 1: Continued.

```
                evaluate children and mutated individuals in the new population with test data and
                ONE/cwk-NN/wk-NN OVA
                add the elite individual without changes into the new population
                //Survivor Selection
                if nr of individuals in the new population is larger than population_size then
                        sort cases descending by their fitness rate
                        discard the last cases in order to have correct nr of individuals in the population
                 else if nr of individuals in the new population is smaller than population_size then
                        select randomly missing cases from the old population
                end if
        end while
        initialize methods with training data D-d_cv_round:
                cwk-NN and wk-NN OVA: HVDM initialization
                ONE: fitness value calculation for values of attributes
        evaluate the individual with the highest fitness rate after GA with testing data d_cv_round and
        ONE/cwk-NN/wk-NN OVA
end for
```

Pseudocode 1: Pseudocode of the genetic algorithm used in the evolution of the attribute weight values with 10-fold cross-validation.

Table 2: Parameters used with the genetic algorithm.

Genetic algorithm parameters	
Crossover rate	0.8
Mutation rate	0.01
Population size	21
Generation	20 (and 100 for ONE)
Elitism	Yes (1 individual)

attribute weighted k-nearest neighbour method using neighbour's class based attribute weighting (*cwk*-NN), and with the attribute weighted k-nearest neighbour method using one-versus-all the other (OVA) classifiers (*wk*-NN OVA). The evaluation methods are defined more accurately in Section 3.2. During the genetic algorithm runs, for each individual in the population its fitness rate was calculated with the method at hand; that is, the individual was evaluated against the method. Within the methods *cwk*-NN and ONE, the fitness rate for the individual was defined with a total classification accuracy (ACC) and within the *wk*-NN OVA with a true positive rate (TPR). The total classification accuracy was used with the ONE and the *cwk*-NN because all seven disease classes were classified at the same time whereas the *wk*-NN OVA concentrated on one disease class (and its weight set) at a time. During *GA wk-NN OVA* runs, it was more important to find the weight set that separated well the cases of the disease class at hand from the others than to classify the other cases also well.

The total classification accuracy showed the percentage of all correctly classified cases within the data set:

$$\mathrm{ACC} = 100\frac{t_{\mathrm{pos}}}{n_{\mathrm{cases}}}\%, \tag{1}$$

where t_{pos} was the total number of cases correctly classified within classes and n_{cases} was the total number of cases used in the classification. The true positive rate expressed the percentage of correctly inferred cases within the class as

$$\mathrm{TPR} = 100\frac{t_{\mathrm{pos}_c}}{n_{\mathrm{cases}_c}}\%, \tag{2}$$

where t_{pos_c} was the number of correctly classified cases in class c and n_{cases_c} was the number of all cases in class c. With the *cwk*-NN and *wk*-NN OVA methods, the classification performance was calculated from the seven nearest neighbour method (7-NN) results and with the ONE from the first diagnosis suggestion (ONE1). However, for disease class benign recurrent vertigo (BRV) with the *wk*-NN OVA method it was necessary to use the TPR of three nearest neighbours (3-NN) as the fitness rate because of the small size of the disease class at hand. Otherwise the TPR for classifying BRV would have always been zero. Nonetheless, if there occurred a situation where TPR of 3-NN was zero with all individuals in the starting population, a new population was created randomly and evaluated. Random new population was created at most ten times and if the TPR did not change during 10 runs, GA run was ended.

A 10-fold cross-validation (CV) [2] was used in evaluating the classification performance of the genetic algorithm. The data was randomly divided into 10 subsets of approximately equal size. The division was made in a stratified manner to ensure that the class distribution of each subset resembled the skewed class distribution of the entire data set. In the beginning, one cross-validation partition (10% of the data) was left aside to test the performance of the found best individual after genetic algorithm run. The nine cross-validation partitions (90%) were used during the training process. In order to calculate the fitness rate for each individual in the population during genetic algorithm runs, the training data was further divided into two parts: six cross-validation parts were used for training and three cross-validation parts were used for testing the current machine learning method used in the fitness rate calculation. Thus, during the genetic algorithm

run 60%–30% data division was used. After the genetic algorithm run, the individual having the highest fitness rate was declared as a result of weight combination and it was then tested with the left aside test data subset. The 10-fold cross-validation was repeated ten times. In total, there were 100 test runs per each evaluation method used in the genetic algorithm. The same cross-validation divisions were used with all the evaluation methods—that is, each method had the same training and testing sets used during the genetic algorithm runs.

3.1. Starting Population. The starting population consisted of 21 individuals. Each individual included seven different attribute weight sets (weights for 94 attributes), one set for each disease class. Instead of selecting the starting individuals at random, we decided to use good "guesses" as a starting point. Therefore, the starting individuals were based on the attribute weights defined by the domain experts (three different weight set versions) and learnt by three machine learning methods (the Scatter method [21–23] and the weighting method of the instance-based learning algorithm IB4 [24] and its variant IB1w). Based on the weight sets defined by the experts and the machine learning methods, two different modifications were created from weight sets with 50% random mutation, thus having 18 weight sets in total. In addition to these, three totally random weight sets were created into the starting population.

The weight values were computed with the machine learning methods from the imputed data set, that is, from the data set where the missing values of attributes were substituted with the class-wise modes of the qualitative and the class-wise medians of the quantitative attributes. In total, 10.1% of the values of attributes were missing in the data set. The imputation was done class-wise on the basis of the whole data prior to data division into training and testing sets. The calculation of the weights was repeated 10 times for each CV training set in the Scatter, IB4, and IB1w methods and the mean weights of the 10 repetitions were used in the classification to handle the randomness in these methods. The weights defined by the application area experts were the same for each CV training set.

The experts' weights were based on three different combinations. The first weight set included the original attribute weights defined by a group of experienced otoneurological physicians for the decision support system ONE made in the 1990s [25]. The second and the third weight sets were defined by two domain specialists during the upgrade process of the decision support system in the 2000s [16].

The Scatter method is normally used for attribute importance evaluation [21–23]. It calculates a scatter value for an attribute that expresses the attributes' power to separate classes in the data set. For attribute weighting purposes, the scatter values were calculated for each attribute in different class versus other classes' situations. In order to use the scatter values as attribute weights, it was necessary to take inverses of scatter values.

The weight calculation method of the IB4 classification method computes attribute weights independently for each class with a simple performance feedback algorithm [24]. The attribute weights of IB4 reflect the relative relevancies of the attributes in the class. The difference between IB4 and its simpler version IB1w is that IB1w saves all processed cases in its class descriptions and does not discard any cases from the class descriptions during runs. Also, the cases with poor classification records are kept in class descriptions with IB1w whereas IB4 discards these cases based on their past performance during classification.

More detailed description of the machine learning methods Scatter, IB4, and IB1w and their use in weight formation will be given in the paper [17].

In order to have different weight sets comparable to each other during the genetic algorithm runs, the attribute weights were normalized into range [0, 1]. The values of each weight set were divided by the highest weight value occurring in the weight calculation method at issue.

3.2. Evaluation Methods

3.2.1. Nearest Pattern Method of ONE. The first method used within the genetic algorithm to evaluate the performance of the individuals in the population was the inference mechanism of the otoneurological decision support system ONE [26]. Its inference mechanism resembles the nearest neighbour methods of pattern recognition. Instead of searching for the nearest case from the training set, it searches for the most fitting class for a new case from its knowledge base.

In the knowledge base of ONE, a pattern is given to each class that corresponds to one vertigo disease. The pattern can be considered a profile of a disease as it describes its related symptoms and signs. Each class in the knowledge base is described with a set of attributes with weight values expressing their significance for the class. In addition, a fitness value for each attribute value is given to describe how it fits the class. The fitness values for attribute values were computed on the basis of the 60% part of training data. Fitness values can have values between 0 and 100. The fitness value 0 means that the attribute value does not fit the class, whereas the fitness value 100 shows that the value fits the class perfectly. The weight values for attributes were given in the population in the GA; thus, the weight values varied from 0 to 1. The greater the weight value is, the more important the attribute is for the class.

The inference mechanism calculates scores for the classes from the weight and fitness values of the attributes. The score $S(c)$ for a class c is calculated in the following way:

$$S(c) = \frac{\sum_{a=1}^{A(c)} x(a)\, w(c,a)\, f(c,a,j)}{\sum_{a=1}^{A(c)} x(a)\, w(c,a)}, \tag{3}$$

where $A(c)$ is the number of the attributes associated with class c, $x(a)$ is 1 if the value of attribute a is known and otherwise 0, $w(c,a)$ is the weight of the attribute a for class c, and $f(c,a,j)$ is the fitness value for the value j of the attribute a for class c [26]. In the case of quantitative attributes, the fitness values are interpolated by using the attribute values in the knowledge base as interpolation points. The fitness values are altered to the range of 0 to 1 during the inference process.

In addition to the score, the minimum and maximum scores are calculated for the classes using the lowest and the highest fitness values for the attributes having missing values.

The classes are ordered primarily by the score and secondarily by the difference of the minimum and maximum score. If the classes have the same score but one class has a smaller difference between the minimum and maximum scores than the others, the class having the smallest difference is placed higher in order. If the classes have the same score and the minimum and maximum score difference, their order is selected randomly. The class having the highest score is referred to as the best diagnosis suggestion.

Some vertigo diseases resemble each other by having a similar kind of symptoms with other diseases during some phase of the disease and, in addition, some patients can actually have two (or more) vertigo diseases present concurrently [27]. Therefore, it is good to check the classification results of ONE with more than one disease suggestion. In the end, the final diagnostic choice must be made by the physician based on the information given on all alternative diseases [27].

3.2.2. Attribute Weighted k-Nearest Neighbour Method Using Neighbour's Class-Based Attribute Weighting. The other method used in the population evaluation was the attribute weighted k-nearest neighbour method using neighbour's class-based attribute weighting (*cwk*-NN). The distance measure of the basic k-nearest neighbour method [1] was expanded to take the attribute weighting into account [6]. Lee et al. [9] used a similar class-dependent attribute weighting with their modified k-nearest neighbour method where different attribute weight sets for different classes were determined with the adaptive-3FW feature weighting method. With our *cwk*-NN the attribute weighting depends on the disease class of the neighbour case. Thus, there ought to be as many attribute weights sets available as there are classes.

The distance measure used with the *cwk*-NN was the Heterogeneous Value Difference Metric (HVDM) [28] expanded with the attribute weighting. HVDM was used because it can handle both qualitative and quantitative attributes in the data set. The attribute weighted HVDM is defined as

$$\text{weighted\_HVDM}\ (x, y) = \sqrt{\sum_{a=1}^{m} w_{ca} d_a(x_a, y_a)^2}, \quad (4)$$

where m is the number of attributes, c is the disease class of the case y, w_{ca} is the weight of the attribute a in class c, and $d_a(x_a, y_a)$ is the distance between the values x_a and y_a for attribute a. The distance function $d_a(x_a, y_a)$ is defined as

$$d_a(x_a, y_a) = \begin{cases} 1, & \text{if } x \text{ or } y \text{ is unknown} \\ \text{normalized\_vdm}_a(x_a, y_a), & \text{if } a \text{ is qualitative} \\ \text{normalized\_diff}_a(x_a, y_a), & \text{otherwise.} \end{cases} \quad (5)$$

Because HVDM computes distances to qualitative and other attributes with different measurement ranges, it is necessary to scale their results into approximately the same range in order to give each attribute a similar influence on the overall distance [28]. The normalized distance to a quantitative attribute is calculated with (6):

$$\text{normalized\_diff}_a(x_a, y_a) = \frac{|x_a - y_a|}{4\sigma_a}, \quad (6)$$

where σ_a is the standard deviation of the numeric values of attribute a in the training set of the current classifier, and to a nominal attribute with (7):

$$\text{normalized\_vdm}_a(x_a, y_a) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2}, \quad (7)$$

where C is the number of output classes in the problem domain (in this case $C = 7$), $N_{a,x(y),c}$ is the number of cases in T that have a value x (or a value y) for attribute a and the output class c, and $N_{a,x(y)}$ is the number of cases in T that have a value x (or a value y) for attribute a [28]. In other words, we are calculating the conditional probabilities to have the output class c when having attribute a with the value x (or the value y).

This approach allowed modifications of all the weights at the same time.

3.2.3. Attribute Weighted k-Nearest Neighbour Method Using One-versus-All Classifiers. In addition to the neighbour's class-based attribute weighting the attribute weighted k-nearest neighbour method was tested with one-versus-all classifiers (*wk*-NN OVA). Within this method, the multiclass classification problem was converted into multiple binary classifiers—that is, the m class problem was divided into m binary problems [29]. Each binary OVA classifier was trained to separate a class from all the other classes by marking the cases of this one class as member cases and the cases of the other classes as nonmember cases in the training set.

The attribute weighted k-NN OVA is an instance-based learning method that searches for the k most similar cases (neighbours) of a new case from each classifier separately. There is one classifier per each class and each classifier gives a vote for the case being a member or nonmember of the class based on the majority class of the k neighbours. The final class of the new case is assigned from a classifier suggesting the case being a member of a class. There can occur a situation in which the new case gets more than one member of a class vote (a tie situation) or all of the classifiers vote for the other class (the case to be a nonmember of all the classes). In a tie situation the class of the new case is determined by searching for the most similar member case from the member voting classifiers. The case gets the class of the member case with the shortest distance to it. When all the classifiers vote for the case to be a nonmember, the basic 1-nearest neighbour classifier using the whole training data containing the original disease classes is employed to find the most similar case (and its class) for the new case.

The distance measure used in the *wk*-NN OVA was also the HVDM measure. The difference in the HVDM description in (4) is that the c is the class of the classifier at issue, not

Table 3: Example evaluation computation time of one population (21 individuals, one generation) in GA runs with different computers.

	Example one population evaluation time			
Computer	*GA ONE*	*GA cwk-NN*	*GA wk-NN OVA*	Specifications
C1	3 min 25 s	48 min 54 s	4 h 57 min 8 s	W7 Intel Core i7-3540M 3.00 GHz, 16 GB RAM
C2	—	49 min 53 s	6 h 59 min 16 s	I3-530 2.93 GHz, 12 GB RAM
C3	—	3 h 47 min 41 s	9 h 41 min 9 s	Q6600 2.4 GHz, 8 GB RAM
C4	—	3 h 12 min 41 s	21 h 14 min 0 s	HP ProLiant DL580 G7 server: 4* Intel Xeon X7560 2.26 GHz, 1 TB RAM
C5	—	3 h 4 min 58 s	7 h 22 min 52 s	DL785 G5 server: 8* AMD Opteron 8360 SE 2.5 GHz, 512 GB RAM
C6	—	—	10 h 34 min 55 s	Intel Core2 Duo E6750 2.66 GHz, 2 GB RAM

Table 4: The ending time of the genetic algorithm runs within different evaluation methods.

Genetic algorithm	*GA ONE*	*GA cwk-NN*	*GA wk-NN OVA*	*GA ONE100*
Ended before 20th generation [%]	75.0	18.0	82.9	39.0*
Ended on 10th generation [%]	48.0	6.0	54.9	12.0*
Ended on 20th generation [%]	25.0	82.0	17.1	61.0*

*The ending generations of the *GA ONE100* runs was examined before 100th generation, on 50th generation and on 100th generation.

the class of the case y. In addition, in (7) *wk*-NN OVA has two output classes ($C = 2$). The data in the learning set T of the classifier is divided into the member and nonmember classes.

4. Results

The results of the GA runs with ONE and *cwk*-NN as an evaluation method were the averages of the 10 times repeated 10-fold cross-validation whereas the results with the *wk*-NN OVA were the averages of the 5 times repeated 10-fold cross-validation. The 10-fold cross-validation was repeated only five times with the *GA wk-NN OVA* due to its huge computation time. For example, the evaluation of a population (21 individuals in one generation in a GA run) in one cross-validation set with the *GA ONE* lasted 3 minutes and 25 seconds, with the *GA cwk-NN* 48 minutes and 54 seconds, and with the *GA wk-NN OVA* 4 hours, 57 minutes, and 8 seconds when running the GA with the computer C1 (Table 3). With the other computers, the computation was even slower. Thus, at worst, the computation time of one cross-validation set lasting 20 generations with the computer C1 and *GA wk-NN OVA* was over four days (over 12 days with C4) assuming that within each generation all individuals were evaluated. In practice, the number of evaluated individuals varied within generations due to the crossover and the mutation. Notice that computers C4 and C5 were servers having several other users simultaneously and, thus, we had only minor part of their CPU in use. During *GA cwk-NN* and *GA wk-NN OVA* runs, the GA was run parallel in five computers, thus, having at best 11 parallel GA runs in process. *GA ONE* was run only with the computer C1.

The number of generations in the GA runs with all used evaluation methods varied from 10 to 20. In total, 75.0%, 18.0%, and 82.9% of GA runs ended before the 20th generation due to having the same best accuracy (*GA ONE* and *GA cwk-NN*) or TPR (*GA wk-NN OVA*) in 10 consecutive GA runs with ONE method, *cwk*-NN, and *wk*-NN OVA, respectively (Table 4). With the *GA wk-NN OVA*, all the GA runs with the disease classes sudden deafness, traumatic vertigo, and benign recurrent vertigo ended before the 20th generation and with the other classes from 58.0% to 88.0% of the runs. If the number of ending generation was 10, this meant that the best ACC or TPR in the population did not change at all during the GA run and, therefore, the run was ended. *GA cwk-NN* ended after 10 generations only in 6.0% of the GA runs whereas *GA ONE* and *GA wk-NN OVA* ended during the GA runs around half of runs (in 48.0% and 54.9% of runs, resp.). In the *GA wk-NN OVA* runs, this happened especially with disease class traumatic vertigo where all CV runs ended after 10 generations and with sudden deafness (96.0%) and benign recurrent vertigo (94.0%). The other disease classes ended during the *GA wk-NN OVA* runs after 10 generations from 12.0% (acoustic neurinoma) to 34.0% (vestibular neuritis) of the runs. Most of the *GA cwk-NN* runs lasted 20 generations (82.0%) whereas only a fourth of the *GA ONE* runs and 17.1% of the *GA wk-NN OVA* runs went through 20 generations.

Within the *GA wk-NN OVA* runs of the disease class benign recurrent vertigo occurred situations where the TPRs in the starting population were zero regardless of using the TPR of 3-NN instead in population evaluation. The TPR of 3-NN was used with BRV instead of 7-NN because of the small size of the disease class. The TPRs of starting individuals were zero in 30 out of 50 cross-validation sets within the *GA wk-NN OVA* run concentrating on the BRV. In this case, new starting individuals were created randomly. Random individual creation was repeated in different cross-validation sets from one to five and nine times. The *GA wk-NN OVA* run ended if the TPR of starting population stayed zero ten times. This happened in 14 (28.0%) cross-validation sets only with the disease class benign recurrent vertigo.

In order to see the effect of genetic algorithm on the population, we examined the worst and the best total classification accuracies of individuals (the attribute weight vectors) in the beginning and in the end of the genetic algorithm run. The mean worst and the mean best total accuracies and their standard deviations with GA runs using ONE and *cwk*-NN as

Table 5: The mean and its standard deviation of the best and worst total classification accuracies of individuals in the starting and ending populations occurring during different GA runs within 10 times (in *GA wk-NN OVA* 5 times) repeated 10-fold cross-validation.

Method	Population	Best accuracy [%]		Worst accuracy [%]	
		Mean	Std dev.	Mean	Std dev.
GA ONE (ONE1)	start	74.0	0.8	49.8	1.6
	end	73.8	0.7	61.4	2.8
	end 100	73.9	0.9	66.5	2.0
GA cwk-NN (7-NN)	start	63.6	1.6	27.9	2.2
	end	68.3	1.9	56.2	2.2
GA wk-NN OVA (7-NN)	start	79.2	0.5	75.3	0.5
	end	78.6	0.9	78.7	0.8

Table 6: The starting point of the genetic algorithm using ONE inference (*GA ONE*), the attribute weighted k-nearest neighbour method with neighbour's class-based attribute weighting (*GA cwk-NN*) and with OVA classifiers (*GA wk-NN OVA*) as evaluation method. The true positive rates (TPR) of seven disease classes and the total classification accuracies of the best individual from the starting population are given in percentages (%) from 10 times (five times with *GA wk-NN OVA*) repeated 10-fold cross-validation.

	Disease	ANE	BPV	MEN	SUD	TRA	VNE	BRV	Median TPR	Total accuracy
	Cases	131	173	350	47	73	157	20		951
GA ONE	ONE1	63.5	55.0	91.1	67.4	84.0	67.5	37.0	67.4	74.0
	ONE12	76.0	84.7	96.6	97.0	96.3	75.4	69.5	84.7	87.5
	ONE123	88.1	94.7	98.1	99.6	99.9	84.6	86.0	94.7	93.8
GA cwk-NN	1-NN	47.6	50.2	75.7	28.7	59.0	55.0	10.5	50.2	58.8
	3-NN	48.9	52.5	82.2	24.0	58.9	57.0	9.0	52.5	61.9
	5-NN	49.0	54.4	85.1	21.1	57.0	56.5	8.5	54.4	62.9
	7-NN	48.9	55.0	86.6	19.6	56.3	57.8	5.5	55.0	63.6
	9-NN	49.2	56.0	87.8	16.4	53.4	57.5	3.5	53.4	63.7
GA wk-NN OVA	1-NN	70.4	73.5	85.0	67.2	62.7	78.2	19.0	70.4	75.8
	3-NN	71.1	75.8	91.8	73.2	61.1	79.4	18.0	73.2	79.2
	5-NN	70.7	75.7	92.8	74.5	62.5	79.5	15.0	74.5	79.6
	7-NN	69.9	74.7	93.0	73.2	60.0	80.1	15.0	73.2	79.2
	9-NN	68.9	73.2	93.2	71.9	58.1	80.5	16.0	71.9	78.7

an evaluation method were calculated from 10 times repeated 10-fold cross-validation and with GA runs using *wk*-NN OVA from 5 times repeated 10-fold cross-validation (Table 5). The mean best accuracies stayed approximately the same with the *GA ONE*, whereas the mean best accuracy increased 4.7% with the *GA cwk-NN* and decreased 0.6% with the *GA wk-NN OVA*. The improvement can be seen from the mean worst classification accuracies: the worst accuracy occurring in the population increased during GA runs, especially with the *GA cwk-NN* (28.3%). With the *GA ONE*, the mean worst accuracy improved 11.6% when using at most 20 generations and 16.7% when using at most 100 generations. With the *GA wk-NN OVA*, the improvement was moderate (3.4%) but one must notice that its mean worst classification accuracy was already over 75% in the starting population, which was better than the mean best accuracies of the other methods.

The more detailed results of the *GA ONE*, the *GA cwk-NN*, and the *GA wk-NN OVA* runs in the beginning and in the end with the best individual occurring in the population are given in Tables 6 and 7. The true positive rates of the disease classes are shown with *GA ONE* for the first (ONE1), the first and second (ONE12), and the first, second, and third (ONE123) diagnosis suggestions of ONE and with *GA cwk-NN* and *GA wk-NN OVA* for one, three, five, seven, and nine nearest neighbours (1-NN–9-NN). During cross-validation runs in GA, the individuals were evaluated by the total classification accuracy of the ONE1 with the *GA ONE* and of the 7-NN with the *GA cwk-NN* and by the true positive rate of the 7-NN with the *GA wk-NN OVA* (except with disease class BRV that used the TPR of 3-NN). The true positive rate was used as a fitness rate with the *GA wk-NN OVA* instead of the total accuracy because it concentrated on classifying one disease class at a time whereas *GA ONE* and *GA cwk-NN* classified all seven disease classes at the same time.

Within 20 generations lasting GA, the best improvement between the start population and the end population was yielded with the *GA cwk-NN* that improved the total classification accuracies and the mean true positive rates when using one to nine nearest neighbours in the classification. Total classification accuracy of the *GA cwk-NN* rose at best 5.1%

Table 7: The end result of the genetic algorithm using ONE inference (*GA ONE*), the attribute weighted k-nearest neighbour method with neighbour's class-based attribute weighting (*GA cwk-NN*) and with OVA classifiers (*GA wk-NN OVA*) as evaluation method in population evaluation after at most 20 generations. The true positive rates (TPR) of seven disease classes and the total classification accuracies of the best individual in the end population are given in percentages (%) from 10 times (five times with *GA wk-NN OVA*) repeated 10-fold cross-validation.

	Disease	ANE	BPV	MEN	SUD	TRA	VNE	BRV	Median TPR	Total accuracy
	Cases	131	173	350	47	73	157	20		951
GA ONE	ONE1	63.5	55.4	90.8	66.2	83.0	68.0	31.5	66.2	73.8
	ONE12	77.0	82.7	96.4	93.6	96.2	76.2	62.0	82.7	87.0
	ONE123	87.6	92.8	98.0	98.5	99.5	84.4	84.5	92.8	93.2
GA cwk-NN	1-NN	70.2	50.0	68.4	30.6	70.0	60.3	15.0	60.3	61.1
	3-NN	70.8	53.9	78.1	27.7	72.5	62.9	14.5	62.9	65.9
	5-NN	70.5	56.1	81.5	23.2	71.9	63.8	12.0	63.8	67.4
	7-NN	69.5	56.6	84.7	21.1	71.0	63.9	8.5	63.9	68.3
	9-NN	69.0	57.5	86.6	18.1	69.7	64.1	6.0	64.1	68.8
GA wk-NN OVA	1-NN	71.5	74.1	84.6	67.2	67.1	77.8	18.0	71.5	76.2
	3-NN	71.6	75.3	91.7	74.9	66.8	78.7	16.0	74.9	79.5
	5-NN	70.4	73.6	92.2	77.0	63.6	79.2	14.0	73.6	79.1
	7-NN	70.4	71.8	92.6	77.0	59.5	79.6	13.0	71.8	78.6
	9-NN	70.5	72.4	92.7	74.9	59.7	79.6	13.0	72.4	78.7

(in 9-NN) and median TPR 10.7% (in 9-NN). The GA had a smaller effect on the results of the *GA ONE* and the *GA wk-NN OVA*. The results in the start population and in the end population stayed quite near each other. Small improvement in the mean total classification accuracy and the mean TPR can be seen with the *GA wk-NN OVA* using one or three nearest neighbours in the classification. Otherwise, the total classification accuracies decreased a bit when using the *GA ONE* and with the *GA wk-NN OVA* using five or seven nearest neighbours in the classification.

Changes within the true positive rates of disease classes compared to the start and end results varied between methods. The *GA cwk-NN* mainly increased the TPRs. During GA runs, it increased the most the TPR of acoustic neurinoma (22.6% in 1-NN) and traumatic vertigo (16.3% in 9-NN). Menière's disease was the only class where the TPR decreased (at worst −7.3% in 1-NN) during *GA cwk-NN* runs. With the *GA ONE*, the TPRs of classes mainly decreased. It decreased the most the TPR of benign recurrent vertigo (−7.5% in ONE12) and sudden deafness (−3.4% in ONE12). However, small increase in TPR can be seen with acoustic neurinoma (1.0% in ONE12) and with vestibular neuritis (0.8% with ONE12). With the *GA wk-NN OVA*, some TPRs increased and some decreased. The TPR increased the most with traumatic vertigo (5.8% in 3-NN) and sudden deafness (3.8% in 7-NN) and decreased the most with benign recurrent vertigo (−3.0% in 9-NN) and benign positional vertigo (−2.9% in 7-NN).

Because the computation time with the ONE method was so much faster than with the k-nearest neighbour methods, the evolution of the population with *GA ONE* runs was tested also with 100 generations in addition to the 20 generations. The ending condition was also changed: the GA run ended if the maximum accuracy stayed the same in 50 successive runs or 100 generations were run. In total, 39.0% of the *GA ONE100* runs ended before the 100th generation and within 12.0% of the runs there was no change in the best total classification accuracy during 50 generations (Table 4). The classification results of the *GA ONE100* runs are given in Table 8. The increase of generations from 20 to 100 did not affect much the mean total classification accuracy nor the mean median TPR. Within disease classes, benign recurrent vertigo suffered the most from the generation increase: its true positive rate decreased at worst −16.0% (ONE12) compared to the starting population and −9.5% (ONE123) compared to the 20th generation. The best TPR increase was achieved with acoustic neurinoma: 3.9% from the starting population and 3.6% from the 20th generation.

Table 8: The end result of the genetic algorithm using ONE inference in population evaluation after at most 100 generations. True positive rates and the total classification accuracies of the best individual in the end populationare given in percentages [%] from 10 times repeated 10-fold cross-validation.

Disease	Cases	*GA ONE 100*		
		ONE1	ONE12	ONE123
ANE	131	67.1	79.9	89.6
BPV	173	56.9	82.0	92.8
MEN	350	89.9	96.1	97.9
SUD	47	61.7	90.9	97.0
TRA	73	80.3	96.4	99.7
VNE	157	69.6	78.7	86.0
BRV	20	23.0	53.5	75.0
Median TPR		67.1	82.0	92.8
Total accuracy	951	73.9	87.3	93.5

5. Discussion and Conclusion

Genetic algorithm runs were done with three different population evaluation methods in order to see whether the classification performance of the attribute weighted methods based on the nearest neighbour search can be improved when using the genetic algorithm in the evolution of attribute weighting. The attribute weighting in the starting population was based on the weights described by the application area experts and machine learning methods instead of random weight setting. The genetic algorithm runs were done separately with the nearest pattern method of ONE (*GA ONE*), with the attribute weighted k-nearest neighbour method using neighbour's class-based attribute weighting (*GA cwk-NN*), and with the attribute weighted k-nearest neighbour method using one-versus-all classifiers (*GA wk-NN OVA*). The 10-fold cross-validation was repeated 10 times with *GA ONE* and *GA cwk-NN* and 5 times with *GA cwk-NN OVA* due to its huge computation time.

The GA runs lasted at maximum 20 generations, 10 generations if there were no change in the best classification accuracy. Most of the GA runs with *GA ONE* and *GA wk-NN OVA* ended before the 20th generation (75.0% and 82.9%, resp.) and around half (!) of the GA runs ended without a change in the best classification (ended after 10 generations; 48.0% and 54.9%, resp.). Only 18.0% of the *GA cwk-NN* runs ended before the 20th round and 6.0% after 10 generations.

The total classification accuracies and the mean true positive rates were improved within *GA cwk-NN* runs whereas with *GA ONE* and *GA wk-NN OVA* the results in the beginning and in the end population stayed quite near each other. One reason why the GA did not improve much the total classification accuracies with the *GA ONE* and the *GA wk-NN OVA* might be that the attribute weights used in the starting population were already optimized for separate disease classes. In addition, also the fitness values for ONE method can be said to be the best occurring fitness values because they were computed from the otoneurological data with the machine learning method.

Hussein et al. [8] noticed that in some applications a strict cost-benefit analysis may rule out the use of genetic algorithm optimization because of its increase in processing time (e.g., 100–150% increase in counting time compared to the basic classifier with 200 train and test cases and over 400% when using 3824 train cases and 1797 test cases with k-NN leave-one-out). Also, Kelly and Davis [6] admit that it can take a tremendous amount of time to find high-performance weight vectors for variably weighted machine learning methods. The results in [3] showed that the extensions of the k-NN yielded generally better results at the cost of speed since all extensions required a training phase. In this research, the *GA wk-NN OVA* was really time-consuming compared to *GA cwk-NN* and *GA ONE*. However, if the weight calculation needs to be done only once or quite seldom, the time issue is not that crucial, especially if it improves the performance of the method.

In this study the weights set by the experts and learnt by machine learning methods were used as a starting point. This helped a lot the search of appropriate weights but there might be different attribute weight combinations with as good or even better classification results. Therefore it would be good to test genetic algorithm also with totally random starting population and with several different parameters in offspring creation and mutation.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The first author acknowledges the support of Onni and Hilja Tuovinen Foundation, Oskar Öflund's Foundation, and Finnish Cultural Foundation, Päijät-Häme Regional fund who granted scholarships for her postgraduate studies. The authors are grateful to Docent E. Kentala, M.D., and Professor I. Pyykkö, M.D., for their help in collecting the otoneurological data and medical advice.

References

[1] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[2] T. Mitchell, *Machine Learning*, McGraw-Hill, New York, NY, USA, 1997.

[3] Z. Voulgaris and G. D. Magoulas, "Extensions of the k nearest neighbour methods for classification problems," in *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications (AIA '08)*, pp. 23–28, ACTA Press, Anaheim, Calif, USA, February 2008.

[4] J. M. Sotoca, J. S. Sánchez, and F. Pla, "Estimating feature weights for distance-based classification," in *Proceedings of the 3rd International Workshop on Pattern Recognition in Information Systems (PRIS '03)*, Angers, France, 2003.

[5] E. Marchiori, A. Ngom, E. Formenti, J.-K. Hao, X.-M. Zhao, and T. van Laarhoven, "Class dependent feature weighting and k-nearest neighbor classification," in *Proceedings of the 8th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB '13)*, vol. 7986 of *LNBI*, pp. 69–78, Springer, Berlin, Germany, 2013.

[6] J. D. Kelly and L. Davis, "A hybrid genetic algorithm for classification," in *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI '91)*, vol. 2, pp. 645–650, Morgan Kaufmann, San Franciso, Calif, USA, 1991.

[7] W. F. Punch, E. D. Goodman, M. Pei, L. Chia-Shun, P. Hovland, and R. Enbody, "Further research on feature selection and classification using genetic algorithms," in *Proceedings of the 5th International Conference on Genetic Algorithms (ICGA '93)*, pp. 557–564, University of Illinois, Champaign, Ill, USA, 1993.

[8] F. Hussein, N. Kharma, and R. Ward, "Genetic algorithms for feature selection and weighting, a review and study," in *Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR '01)*, pp. 1240–1244, Seattle, Wash, USA, 2001.

[9] H. Lee, E. Kim, and M. Park, "A genetic feature weighting scheme for pattern recognition," *Integrated Computer-Aided Engineering*, vol. 14, no. 2, pp. 161–171, 2007.

[10] D. Mateos-García, J. García-Gutiérrez, and J. C. Riquelme-Santos, "Label dependent evolutionary feature weighting for remote sensing data," in *Proceedings of the 5th International Conference on Hybrid Artificial Intelligence Systems*, pp. 272–279, Springer, 2010.

[11] D. Mateos-García, J. García-Gutiérrez, and J. C. Riquelme-Santos, "On the evolutionary optimization of k-NN by label-dependent feature weighting," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2232–2238, 2012.

[12] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, Mass, USA, 1989.

[13] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, Mass, USA, 1996.

[14] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, Berlin, Germany, 1992.

[15] A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*, Springer, Berlin, Germany, 2003.

[16] K. Varpa, K. Iltanen, M. Juhola et al., "Refinement of the otoneurological decision support system and its knowledge acquisition process," in *Proceedings of the 20th International Congress of the European Federation for Medical Informatics (MIE '06)*, pp. 197–202, Maastricht, The Netherlands, 2006.

[17] K. Varpa, K. Iltanen, M. Siermala, and M. Juhola, "Attribute weighting with scatter and instance-based learning methods evaluated with otoneurological data," *International Journal of Computational Medicine and Healthcare*, 2013.

[18] E. Kentala, I. Pyykkö, Y. Auramo, and M. Juhola, "Database for vertigo," *Otolaryngology—Head and Neck Surgery*, vol. 112, no. 3, pp. 383–390, 1995.

[19] K. Viikki, *Machine learning on otoneurological data: decision trees for vertigo diseases [Ph.D. thesis]*, Department of Computer Sciences, University of Tampere, Tampere, Finland, 2002, http://urn.fi/urn:isbn:951-44-5390-5.

[20] K. A. De Jong, *Analysis of the behaviour of a class of genetic adaptive systems [Ph.D. thesis]*, Computer and Communication Sciences Department, The University of Michigan, Ann Arbor, Mich, USA, 1975, http://hdl.handle.net/2027.42/4507.

[21] M. Siermala, M. Juhola, J. Laurikkala, K. Iltanen, E. Kentala, and I. Pyykkö, "Evaluation and classification of otoneurological data with new data analysis methods based on machine learning," *Information Sciences*, vol. 177, no. 9, pp. 1963–1976, 2007.

[22] M. Juhola and M. Siermala, "A scatter method for data and variable importance evaluation," *Integrated Computer-Aided Engineering*, vol. 19, no. 2, pp. 137–139, 2012.

[23] M. Juhola and M. Siermala, "Scatter Counter program and its instructions," 2014, http://www.uta.fi/sis/cis/research_groups/darg/publications/scatterCounter_2_7_eng.pdf.

[24] D. W. Aha, "Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms," *International Journal of Man-Machine Studies*, vol. 36, no. 2, pp. 267–287, 1992.

[25] E. Kentala, Y. Auramo, M. Juhola, and I. Pyykkö, "Comparison between diagnoses of human experts and a neurotologic expert system," *Annals of Otology, Rhinology and Laryngology*, vol. 107, no. 2, pp. 135–140, 1998.

[26] Y. Auramo and M. Juhola, "Modifying an expert system construction to pattern recognition solution," *Artificial Intelligence in Medicine*, vol. 8, no. 1, pp. 15–21, 1996.

[27] E. Kentala, Y. Auramo, I. Pyykkö, and M. Juhola, "Otoneurological expert system," *Annals of Otology, Rhinology and Laryngology*, vol. 105, no. 8, pp. 654–658, 1996.

[28] R. D. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *Journal of Artificial Intelligence Research*, vol. 6, pp. 1–34, 1997.

[29] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011.

PUBLICATION V

Applying One-vs-One and One-vs-All Classifiers in *k*-Nearest Neighbour Method and Support Vector Machines to an Otoneurological Multi-Class Problem

Kirsi Varpa, Henry Joutsijoki, Kati Iltanen and Martti Juhola

In: Moen A *et al.* (eds.), *Studies in Health Technology and Informatics vol. 169, 2011: User Centred Networked Health Care – Proceedings of 23rd International Conference of the European Federation for Medical Informatics (MIE 2011)*, Oslo, Norway, IOS Press, 2011, pp. 579–583
https://doi.org/10.3233/978-1-60750-806-9-579

User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-579

Applying One-vs-One and One-vs-All Classifiers in *k*-Nearest Neighbour Method and Support Vector Machines to an Otoneurological Multi-Class Problem

Kirsi VARPA [a1], Henry JOUTSIJOKI [a], Kati ILTANEN [a] Martti JUHOLA [a]
[a] *Computer Science, School of Information Sciences, University of Tampere, Finland*

Abstract. We studied how the splitting of a multi-class classification problem into multiple binary classification tasks, like One-vs-One (OVO) and One-vs-All (OVA), affects the predictive accuracy of disease classes. Classifiers were tested with an otoneurological data using 10-fold cross-validation 10 times with *k*-Nearest Neighbour (*k*-NN) method and Support Vector Machines (SVM). The results showed that the use of multiple binary classifiers improves the classification accuracies of disease classes compared to one multi-class classifier. In general, OVO classifiers worked out better with this data than OVA classifiers. Especially, the OVO with *k*-NN yielded the highest total classification accuracies.

Keywords. multi-class classification, binary classifiers, otoneurology, *k*-nearest neighbour method, support vector machines

1. Introduction

Multi-class classification problems can be difficult to understand. Especially, if the application domain is not so familiar before, it can be hard to conceptualize the domain. Whenever creating new computer systems into new domains, it is important to have understanding about domain concepts, their relationships and differences. In order to distinguish classes better, one way is to convert the multi-class problem into multiple two-class problems [1, 2]. This may also help separation of classes. Earlier we have studied otoneurological data, for example, by using machine learning (ML) methods like decision trees [3] and neural networks [4]. Previous studies have shown that certain disease classes are difficult to recognize: they easily mix up with other classes [5]. From the literature, studies can be found where this kind of problem has been eased with using One-vs-One (OVO, also called round robin or pairwise class binarization) [6] and One-vs-All (OVA, also known as one-against-all, one-vs-rest) [7] solutions (i.e. using several binary classifiers instead of trying to classify all the classes at the same time with one classifier). Beforehand, it is not possible to say which of these solutions is better than others. Therefore, we examine the use of multiple binary classifiers to help the classification of vertigo data, and to find out which classifier solution seems to work the best with this data.

[1] Corresponding author: Kirsi Varpa, Computer Science, School of Information Sciences, FI-33014 University of Tampere, Finland; E-mail: Kirsi.Varpa@cs.uta.fi.

In this paper, we examine the effect of using multiple binary classifiers instead of using only one multi-class classifier. Binary classifiers used are OVO and OVA classifiers with a *k*-Nearest Neighbour (*k*-NN) method [8] and Support Vector Machines (SVM) [9].

2. Data and Methods

The *k*-NN classifier is a widely used, basic instance-based learning method that searches for the *k* most similar cases of a test case from the training data [8]. It can be used with both binary and multi-class problems. The *k*-NN classifier used in this research was implemented in Java. The nearest cases were searched with *k*= 1, 3, 5, 7, 9, 11 and 13. The best *k*-NN varied between classes, so, we selected NN classifier with *k*=5 (5-NN) into the comparison to SVM. (In addition, 5-NN was used in our earlier study [5]). The *k*-NN method used Heterogeneous Value Difference Metric (HVDM) [10] since our data included nominal, ordinal and quantitative attributes.

SVM is a newer, more sophisticated ML method to be used in the separation between two classes [9]. It is a kernel-based classification method [11, 12]. Originally, it was made for the binary classification tasks, but later it has been extended for the multi-class cases [13]. The basic idea in SVM is to generate an input space dividing hyperplane such that the margin, the distance between the closest members of both classes, is maximized. The use of SVM was expanded by the invention of kernel trick, where the input space is mapped with a non-linear transformation into higher dimensional space [14, 15]. In the research, we used the binary SVM implementation of Bioinformatics Toolbox of Matlab with the Least-Square method [16] as a basis for the multi-class extensions. SVM runs were made with linear, polynomial (d=2,3,4,5), Multilayer Perceptron (MLP) (scale κ in [0.2,10]; bias δ in [-10,-0.2]) and Gaussian Radial Basis Function (RBF) (scaling factor σ in [0.2,10]) kernels with box constraints [0.2,10] (κ, δ and σ with intervals 0.2). The best kernel functions, linear and RBF, were selected into comparison.

ML methods were tested with an otoneurological data containing 1,030 vertigo cases from nine different vertigo diseases (Table 1). Data was collected at Helsinki University Central Hospital during several years [3]. The dataset used in this research consists of 94 attributes concerning a patient's health status: occurring symptoms, medical history and findings in otoneurologic, audiologic and imaging tests. More detailed information about the collected patient's information is provided in [17] and in [4] 38 main attributes are described. From the 94 attributes, 17 were quantitative (integer or real) and 77 were qualitative: 54 binary (yes/no) and 23 categorical attributes.

Clinical tests are not done to every patient and, therefore, values are missing in several test results. In total, the data had about 11% missing values, which allowed using imputation. Imputation was needed due to calculation of the SVM method. Missing values of qualitative attributes were imputed (substituted) with class modes and missing values of other attributes with class medians. The imputed data was used with *k*-NN in order to keep it comparable to SVM. A 10-fold cross-validation (CV) was repeated 10 times using each time different random data divisions. Training and test set divisions into 10-fold CV were created with Matlab. In divisions, the ratios of disease classes were maintained in different CV folds. CV was used with both ML methods.

In OVA runs, we had nine (*n_classes*) binary classifiers: each one of them was trained to separate one class from the rest. A test sample was input to each classifier and a final class for the test sample was assigned according to the winner-takes-all rule from a classifier suggesting a class. For OVO runs, we trained 36 ($n\_classes \cdot (n\_classes - 1)/2$) binary classifiers between all pairs of the classes. A test sample was solved with each binary classifier.

Table 1. Nine disease classes and their absolute and relative frequencies in the otoneurological data. Average true positive rates (TPR) of disease classes, median of TPR and total classification accuracies with machine learning methods 5-NN and SVM linear and RBF using OVO and OVA classifiers from ten 10-fold cross-validation runs in percents. Used kernel parameters with SVM linear and RBF presented below the table.

			OVO Classifiers			OVA Classifiers		
Disease Name (Abbreviation)	**Cases** 1,030(100%)	**5-NN**	**5-NN**	**SVM linear**	**SVM RBF**	**5-NN**	**SVM linear**	**SVM RBF**
Acoustic Neurinoma (ANE)	131 (12.7)	89.5	95.0	91.6	87.2	90.2	90.6	90.7
Benign Positional Vertigo (BPV)	173 (16.8)	77.9	79.0	70.0	67.0	77.6	73.5	78.6
Menière's Disease (MEN)	350 (34.0)	92.4	93.1	83.8	90.1	89.8	87.8	91.5
Sudden Deafness (SUD)	47 (4.6)	77.4	94.3	88.3	79.4	87.4	61.3	58.1
Traumatic Vertigo (TRA)	73 (7.1)	89.6	96.2	99.9	99.3	77.7	79.9	96.7
Vestibular Neuritis (VNE)	157 (15.2)	87.7	88.2	82.4	81.4	85.0	85.4	84.3
Benign Recurrent Vertigo (BRV)	20 (1.9)	3.0	4.0	20.0	16.5	8.0	21.0	8.0
Vestibulopatia (VES)	55 (5.3)	9.6	14.0	16.5	22.8	15.8	15.3	13.5
Central Lesion (CL)	24 (2.3)	5.0	2.1	26.0	28.5	15.0	19.0	15.8
Median of TPR		77.9	88.2	82.4	79.4	77.7	73.5	78.6
Total Classification Accuracy		79.8	82.4	77.4	78.2	78.8	76.8	79.4

Linear kernel with box constraint $bc = 0.20$ (OVO and OVA)
RBF kernel with $bc = 0.4$ and $\sigma = 8.20$ (OVO), $bc = 1.4$ and $\sigma = 10.0$ (OVA)

In OVO, the results of pairwise decisions were combined, thus having 36 class suggestions (votes) for the class of the test sample altogether. The final class for the test sample was chosen by the majority voting method, the max-wins rule [1]. A class, which gained the most votes, was chosen as the final class.

If a tie situation occurred in the max-wins (OVO) or winner-takes-all (OVA) rules, the final class, within the tied classes, was solved in SVM by 1-NN, whereas *k*-NN searched for the nearest case from the classifiers belonging to the tied classes and selected the class with minimum distance to the test case. If the test case did not get any class by using *k*-NN with OVA (every classifier voted 0), the class was searched from the whole learning set with normal 1-NN.

3. Results

In the Table 1, mean true positive rates (TPRs) and total classification accuracies of the ten 10-fold cross-validations are presented for 5-NN and SVM with linear and RDF

kernels. Both methods were run by using OVO and OVA classifiers. The 5-NN method was also run in a basic way by using all nine disease classes in a classifier, i.e. all of the training cases class labels were used when searching for the nearest case to the test sample. The basic 5-NN was used as a baseline in the comparison of the predictive accuracies of the methods.

The mean number of tie situations occurring during 10 times repeated 10-fold CV with OVO classifiers was 20.3 with 5-NN (standard deviation SD=4.8), 7.2 using SVM linear (SD=2.6) and 2.6 with SVM RBF (SD=1.5). With OVA classifiers, the number of ties was higher, as expected: 5-NN 167.1 (SD=3.8), SVM linear 49.8 (SD=4.7) and SVM RBF 25.8 (SD=4.2). With 5-NN OVA classifier all of the ties (16.2%) happened when a case could not be classified at all, ties with all nine classifiers, whereas 5-NN OVO classifier had ties (2.0%) with two or three classes (mainly BPV, MEN and VES).

The results show that the use of multiple binary classifiers improves the TPRs of disease classes. The best results were yielded with OVO in 5-NN: it had the highest median of TPR and total accuracy. With this data, the OVO classifiers mainly increase TPRs and the total accuracies, whereas OVA classifiers have slightly decreasing effect on classification. However, there were exceptions also with this: SVM with MLP and polynomials 4 and 5 worked better with OVA classifiers. Usually, MLP is one of the best kernel functions used in SVM, but with this data it did not work at all (total accuracy 25.5% with OVO and 68.5% with OVA). It could also be seen with *k*-NN that the bigger *k*, the closer the results with OVO, OVA and the basic *k*-NN came (except with disease classes SUD and TRA).

4. Discussion

In this research, we concentrated on studying the effect of splitting the multi-class problem into several binary classifiers and the voting procedure within two different ML methods, the *k*-NN and SVM classifiers. Splitting a problem into several binary problems helps to understand data better, especially with OVO classifiers in *k*-NN. The OVO classifiers aid to see which classes are difficult to separate and which ones distinguish well from the others.

Diagnosis of the otoneurological disorders is demanding. For example, in [18], 1,167 patients participated in research but only for 872 patients could be made confirmed diagnosis and in [19], ten of the 33 test cases had to be excluded from the test because even the expert physician could not give them a definite diagnosis. Diseases can simulate each other in the beginning having symptoms of similar kind and symptoms can vary in time making recognition difficult [18, 20]. Classification accuracy of the medical professionals with the data of this study having 1,030 cases has not been tested because this would be an enormous task for them to do. However, a smaller number of cases (23) have been classified with a group of physicians [19].

We need to remember that classification tasks in this research were performed with the imputed data. In real life, there usually occur missing data because clinical tests are not done to every patient automatically. Thus, TPRs and total classification accuracies in this research might be a little bit higher than with the original data having missing values.

There occur some differences in the way how ML methods used in the research handle data. SVM treats each attribute as quantitative, whereas *k*-NN using HVDM

distance metric makes a different calculation depending on the type of the attribute (quantitative or qualitative).

In the future, we shall expand the use of the voting procedure to involve handling the results of several different classification methods (e.g. *k*-NN, nearest pattern method of an otoneurological expert system [21] and Naive Bayes [8]), thus, forming a hybrid decision support aid. Being able to use results of several ML methods simultaneously strengthens the support of decision making.

Acknowledgements: The authors wish to thank Erna Kentala, M.D., and prof. Ilmari Pyykkö, M.D., for their help in data collection during the years and their valuable aid in domain expertise. The first and second authors acknowledge the support of the Tampere Doctoral Programme in Information Science and Engineering (TISE).

References

[1] Friedman JH. *Another approach to polychotomous classification.* Stanford University; 1996 Oct.14 p.
[2] Allwein EL, Schapire RE, Singer Y. Reducing multiclass to binary: a unifying approach for margin classifiers. *J Mach Learn Res.* 2000;1:113–141.
[3] Viikki K. *Machine learning on otoneurological data: decision trees for vertigo diseases [PhD Thesis]*. Tampere, Finland: University of Tampere; 2002.
[4] Siermala M, Juhola M, Kentala E. Neural network classification of otoneurological data and its visualization. *Comput Biol Med*. 2008;38(8):856–866. doi:10.1016/j.compbiomed.2008.05.002.
[5] Varpa K, Iltanen K, Juhola M. Machine learning method for knowledge discovery experimented with otoneurological data. *Comput Methods Programs Biomed*. 2008;91(2):154–164. doi:10.1016/j.cmpb.2008.03.003.
[6] Fürnkranz J. Round robin rule learning, In: Brodley CE, Danyluk AP, editors. ICML-01. *Proceedings of the 18th International Conference on Machine Learning*; 2001. Williamstown, MA: Morgan Kauffman; 2001. P.146–153.
[7] Rifkin R, Klautau A. In defense of one-vs-all classification. J Mach Learn Res. 2004;5:101–141.
[8] Mitchell T. *Machine Learning.* New York: McGraw-Hill;1997.
[9] Debnath R, Takahide N, Takahashi H. A decision based one-against-one method for multi-class support vector machine. *Pattern Anal Appl*. 2004;7(2):164–175. doi:10.1007/s10044-004-0213-6.
[10] Wilson DR, Martinez TR. Improved heterogeneous distance functions. *J Artif Intell Res.* 1997;6:1–34.
[11] Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20:273–297.
[12] Vapnik VN. *The Nature of Statistical Learning Theory*. 2nd ed. Springer; 2000.
[13] Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw*. 2002;13(2):415–425.
[14] Christiani N, Shawe-Taylor J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press; 2003.
[15] Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov.* 1998;2:121–167.
[16] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. *Neural Processing Letters.* 1999;9:293–300.
[17] Kentala E, Pyykkö I, Auramo Y, Juhola M. Database for vertigo. *Otolaryngol Head Neck Surg*. 1995;112:383–390.
[18] Kentala E. Characteristics of six otologic diseases involving vertigo. *Am J Otol.* 1996;17(6):883–892.
[19] Kentala E, Auramo Y, Juhola M, Pyykkö I. Comparison between diagnoses of human experts and a neurotologic expert system. *Ann Otol Rhinol Laryngol 1998*;107(2):135–140.
[20] Havia M. *Menière's disease prevalence and clinical picture [PhD Thesis]*. Helsinki: Department of Otorhinolaryngology, University of Helsinki; 2004.
[21] Auramo Y, Juhola M. Comparison of inference results of two otoneurological expert systems. *Int J Biomed Comput.* 1995;39:327–335.