This is the post print version of the article, which has been published in International Journal of Data Science, 2017, 2 (3), 173-204. http://dx.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 utilized in weighting the attributes. The machine learnt weighting schemes, 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. (2015) 'Attribute weighting with Scatter and instance-based learning methods evaluated with otoneurological data', *Int. J. Data Science*, Vol. X, No. X, pp.X–X.

Biographical notes: Kirsi Varpa received her MSc in 2005 in Computer Science 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 degree in 1997 from the University of Kuopio and PhD degree 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 utilize 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 organized 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 utilized 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 utilized 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 utilized 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 optimization 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, like 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 utilized 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 utilized 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 1,030 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

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

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

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

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 upgrading of the decision support system (Varpa et al., 2006), when new attributes were added to the knowledge base.

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 as 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 utilized 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 utilized 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.

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

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

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

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 *i*th 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

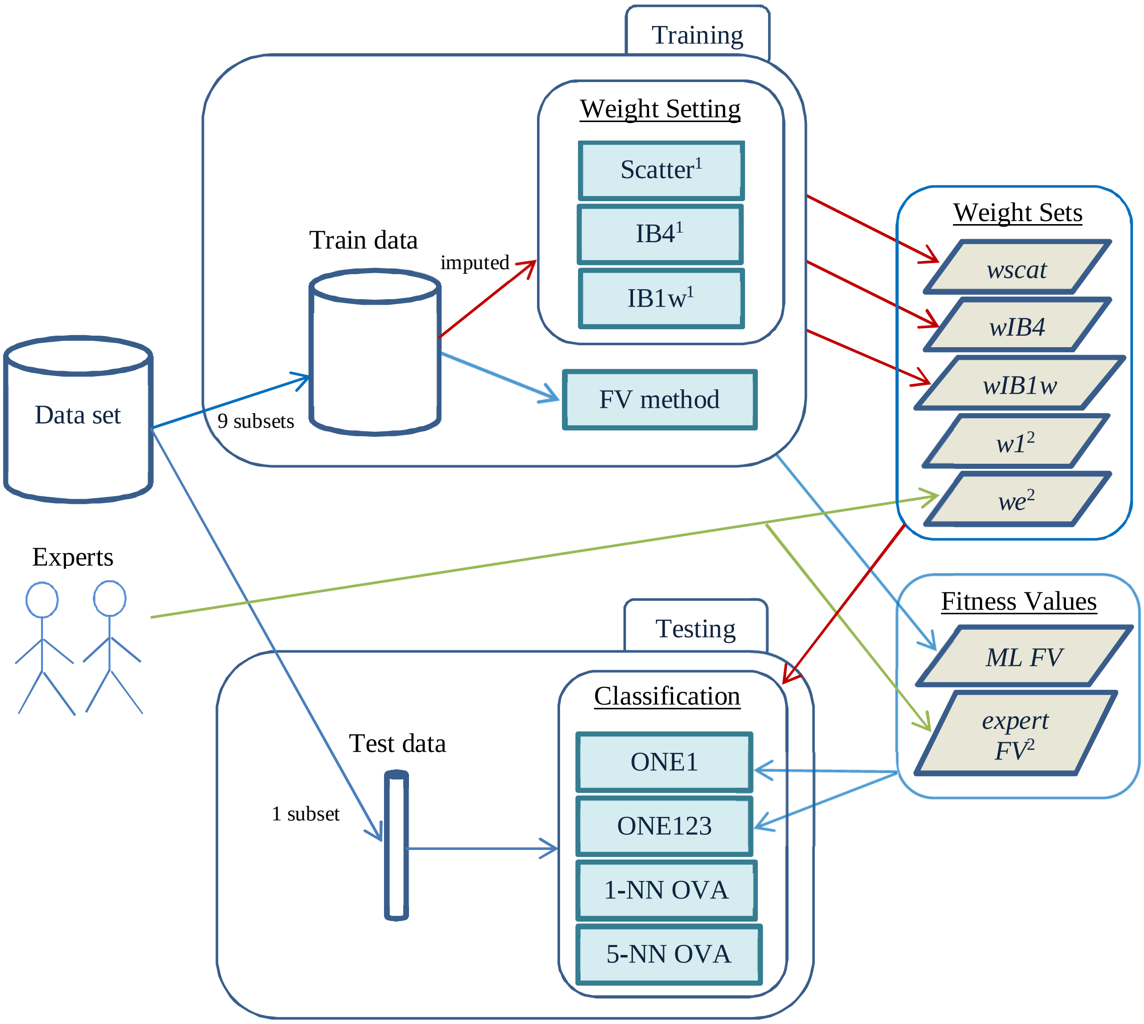
We used a 10-fold cross-validation (CV) (Mitchell, 1997) to evaluate the classification performance of the ONE and the attribute weighted *k*-nearest neighbour OVA methods combined with different weighting schemes. In the 10-fold cross-validation, the data was randomly divided into 10 subsets of approximately 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:

$$TPR = 100\frac{t_{pos_c}}{n_{cases_c}}\%, \qquad (9)$$

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:

$$ACC = 100\frac{t_{pos}}{n_{cases}}\%, \qquad (10)$$

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):

$$K = \frac{P_o - P_c}{1 - P_c}, \qquad (11)$$

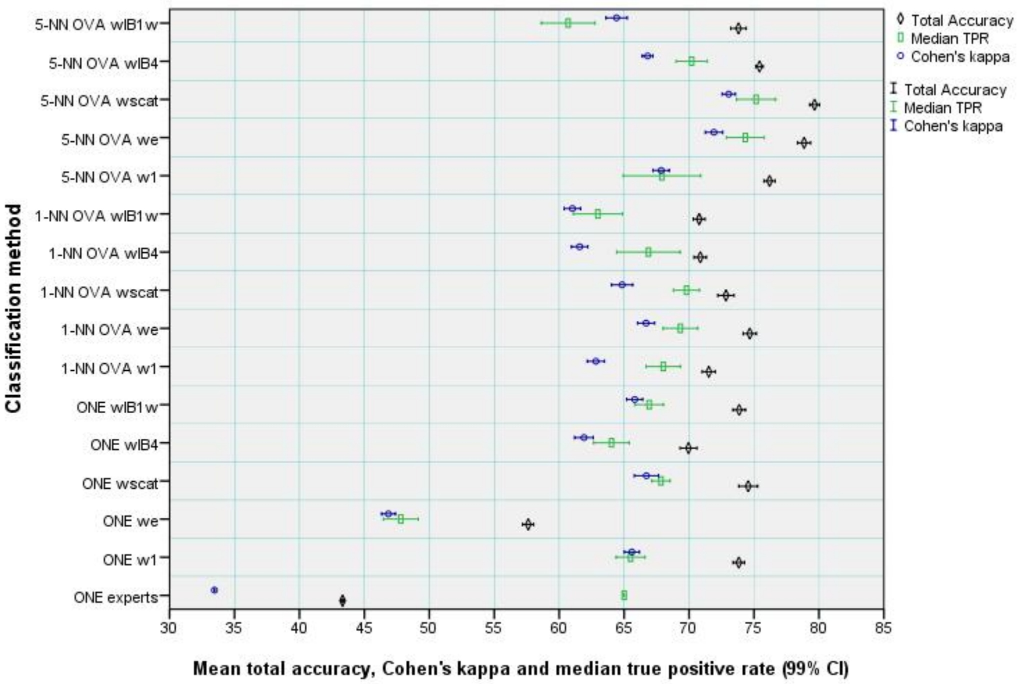
where P_o is the total agreement probability (*i.e.* accuracy) and

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).

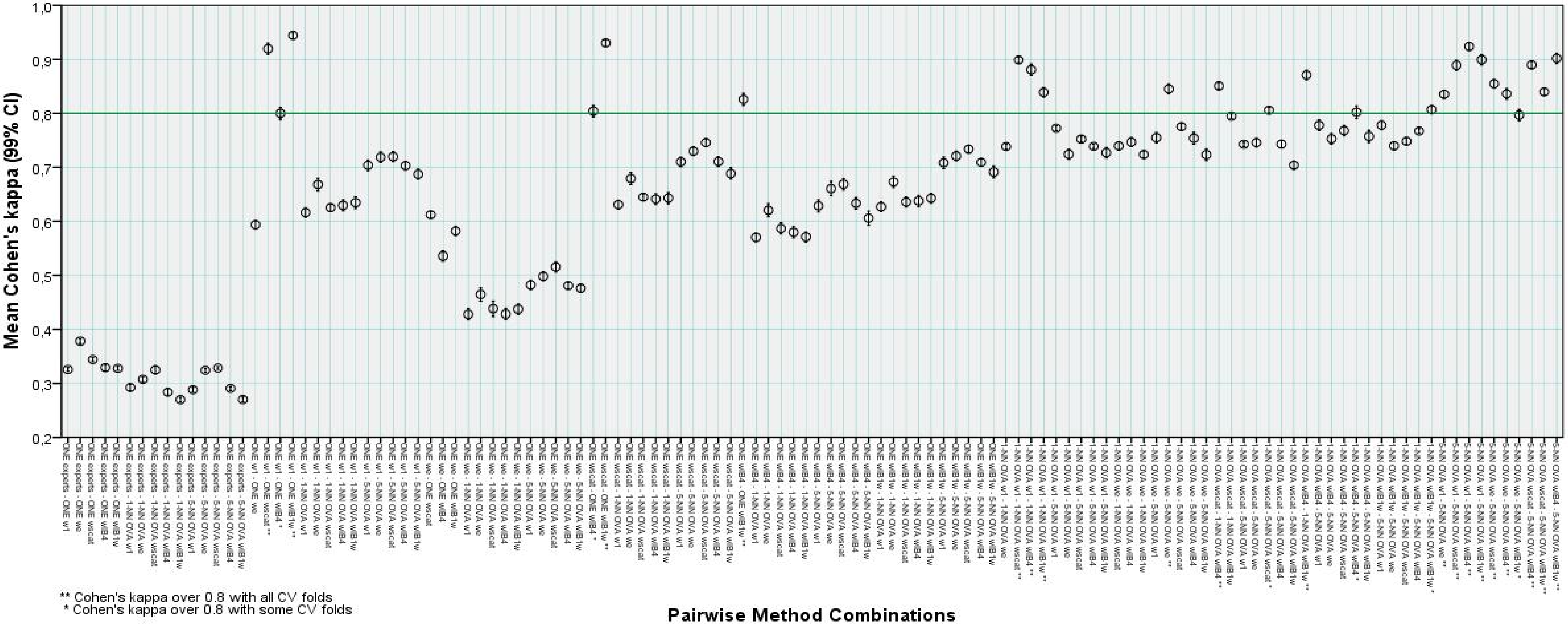


Figure 4 Error bars (with 99% confidence intervals) for the mean Cohen's kappa for pair-wise method combinations from 10 times repeated 10-fold cross-validation with seven disease classes.

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*	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*	max *n*	1-NN OVA we min *n*	max *n*	1-NN OVA wscat min *n*	max *n*	1-NN OVA wIB4 min *n*	max *n*	1-NN OVA wIB1w min *n*	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*	max *n*	5-NN OVA we min *n*	max *n*	5-NN OVA wscat min *n*	max *n*	5-NN OVA wIB4 min *n*	max *n*	5-NN OVA wIB1w min *n*	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		ONE1 wscat	ONE1 wIB4	ONE1 wIB1w
	min *n*	max *n*			
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 min *n*	1-NN OVA w1 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	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 min *n*	5-NN OVA w1 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	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.

The 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 utilized 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 optimization, 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.