Jorma Laurikkala

Knowledge Discovery for Female Urinary Incontinence Expert System

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of
Information Sciences of the University of Tampere, for public discussion in
the Paavo Koli Auditorium of the University on October 26th, 2001, at 12 noon.

DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES
UNIVERSITY OF TAMPERE

A-2001-6

TAMPERE 2001

Supervisor: Professor Martti Juhola, Ph.D.,
Department of Computer and Information Sciences,
University of Tampere,
Finland

Opponent: Professor Heikki Mannila, Ph.D.,
Laboratory of Computer and Information Science,
Helsinki University of Technology,
Finland

Reviewers: Jari Forsström, M.D., Ph.D., docent,
Faculty of Medicine,
University of Turku,
Finland

Tapio Grönfors, Ph.D.,
Cognitive Neurobiology Laboratory,
A. I. Virtanen Institute,
University of Kuopio,
Finland

Department of Computer and Information Sciences
FIN-33014 UNIVERSITY OF TAMPERE
Finland

Electronic dissertation
Acta Electronica Universitatis Tamperensis 137
ISBN 951-44-5193-7
ISSN 145-594X
http://acta.uta.fi

ISBN 951-44-5156-2
ISSN 1457-2060

Tampereen yliopistopaino Oy
Tampere 2001

Abstract

This study addresses the construction of an expert system for the differential diagnosis of female urinary incontinence by using data mining techniques. The motivation for the work was the problematic diagnostic task, and the need for an alternative to the slow and expensive manual knowledge acquisition process in which a knowledge engineer interviews an expert repeatedly to acquire the expert's knowledge. Therefore, the main aims were to develop a decision support tool for the physicians, and to investigate whether data mining techniques could be used to discover diagnostic knowledge automatically from the patient data for the expert system. In this context, special attention was paid to pre-processing of the data and the machine learning methods were researched. This work produced a new machine learning program, Galactica, which is based on genetic algorithms, and the neighbourhood cleaning rule (NCL) that balances the imbalanced class distribution by using an instance-based approach. Comparison of Galactica with different classification methods showed that genetic algorithms were a competitive method for constructing classifiers from medical data. NCL enabled improved identification of difficult small classes, while keeping the classification ability of the other classes at an acceptable level. Galactica, NCL, and other data mining techniques were applied to overcome difficulties with real world data, and to build ability for classification and critique for the incontinence expert system. The study showed that it is possible to develop an expert system with data mining techniques. In 'laboratory conditions' the first version of the system (IES1) correctly classified 94% and 91% of the two batches of the test data, and the medians of the true positive and true negative rates were 97% and 94% in the first test set, and 96% and 90% in the second test set. The expert system was implemented as an Internet-based application that a physician can use with a World Wide Web browser. IES1 seems to be a useful aid for the physicians, but only real world diagnostic work-up will prove its utility in the diagnosis of incontinent women.

Keywords: artificial intelligence, decision support systems, expert systems, medical diagnostic computing, patient diagnosis, data mining, machine learning, genetic algorithms, data pre-processing

Acknowledgments

This work was carried out in the Department of Computer Science and Applied Mathematics, University of Kuopio in the year 1997, and in the Department of Computer and Information Sciences, University of Tampere, during the years 1998-2001.

I wish to thank Professor Martti Juhola, Ph.D., and Professor Seppo Lammi, Ph.D., the former Heads of the Department of Computer Science and Applied Mathematics, for the chance to work in the University of Kuopio. I am grateful to the Heads of the Department of Computer and Information Sciences, Professor Kari-Jouko Räihä, Ph.D., Professor Pertti Järvinen, Ph.D., and Professor Seppo Visala, Ph.D., for the opportunity to continue and complete my work in the University of Tampere.

I wish to express my deepest gratitude to Professor Martti Juhola for giving me a chance to start my post-graduate studies which finally produced this thesis. I am indebted to him for his guidance, his encouragement during the difficulties, and for his patience all the times I had to adjust my schedule.

I also express my gratitude to my other co-authors Jorma Penttinen, M.D., Ph.D., Pauliina Aukee, M.D., Seppo Lammi, Ph.D., and Kati Viikki, M.Sc., for their help in medical, computer scientific, statistical, and machine learning issues. I am grateful to the co-authors of the papers which were not included, but are referred to, in the thesis. In particular, Professor Ilmari Pyykkö, M.D., Ph.D., and Erna Kentala, M.D., Ph.D., have kindly devoted their time for various fruitful discussions.

I have had an opportunity to work with many people both in Kuopio and Tampere. Especially, I wish to thank Ilkka Klemola, M.Sc., Mauno Rönkkö, M.Sc., Vesa Niemilä, M.Sc., Jouni Mykkänen, M.Sc., Timo Tossavainen, M.Sc., and Markku Siermala, M.Sc., with whom I have worked and, occasionally, engaged in inspiring conversations. Kati Viikki, M.Sc., has offered invaluable help during my work - I wish to express to her my most sincere thanks.

I thank Erkki Pesonen, Ph.D., and Erkki Mäkinen, Ph.D., for discussions involving research, and further thanks go to Erkki Mäkinen for being an excellent adversary in the extracurricular floorball games.

I wish to thank Heikki Aalto, Ph.D., Teppo Kuusisto, M.Sc., and the staff of the Department of Obstetrics and Gynaecology of Kuopio University Hospital for their technical assistance. Ilmari Pyykkö, Erna Kentala, Erkki Pesonen, and Pekka Honkavaara, M.D., Ph.D., have provided additional medical data for the study, which is gratefully acknowledged. Virginia Mattila, M.A., has revised the English language of this study.

I am grateful to the reviewers of the thesis, Docent Jari Forsström, M.D., Ph.D., and Professor Tapio Grönfors, Ph.D., for their constructive and prompt comments on the manuscript.

I also wish to thank Leila Tiihonen, Tuula Moisio, and Marja Liisa Nurmi for their patience and assistance in my everlasting struggle with bureaucracy.

Finally I owe my profoundest thanks to my parents and my siblings for all the support they have given me during my studies.

This work was supported financially by the Tampere Graduate School in Information Science and Engineering (TISE), the Oskar Öflund Foundation, the Ella and Georg Ehrnrooth's Foundation, the Savo High Technology Foundation, the Jenny and Antti Wihuri Foundation, the Technology Development Centre (TEKES), the Academy of Finland, Medisinska Forskningsrådet (Sweden), and Mega Elektroniikka Oy.

Tampere, June 2001

Jorma Laurikkala

List of abbreviations

Abbreviation	Description
ANN	Artificial neural network
DSS	Decision support system
EIE	Empirical induction from examples
ENN	Edited nearest neighbour rule
GBML	Genetics-based machine learning
HEOM	Heterogeneous Euclidean-overlap metric
HVDM	Heterogeneous value difference metric
IES1	Incontinence expert system, version 1
KBS	Knowledge-based system
KDD	Knowledge discovery in databases
K-NN	*K*-nearest neighbour
NCL	Neighbourhood cleaning rule
OSS	One-sided selection
RBC	Random bit climber
ROC	Receiver operating characteristics
TNR	True negative rate
TPR	True positive rate
WWW	World wide web

List of the original publications

This thesis is based on the articles referred to in the text by their Roman numerals:

I. Laurikkala J, Juhola M, Lammi S, Penttinen J, Aukee P: Analysis of the imputed female urinary incontinence data for the evaluation of expert system parameters, *Computers in Biology and Medicine* 31 (2001) 239-257.

II. Laurikkala J, Juhola M: A genetic-based machine learning system to discover the diagnostic rules for female urinary incontinence, *Computer Methods and Programs in Biomedicine* 55 (1998) 217-228.

III. Laurikkala J, Juhola M, Lammi S, Viikki K: Comparison of genetic algorithms and other classification methods in the diagnosis of female urinary incontinence, *Methods of Information in Medicine* 38 (1999) 125-131.

IV. Laurikkala J, Juhola M: Nearest neighbour classification with heterogeneous proximity functions. In: Hasman A, Blobel B, Dudeck J, Engelbrecht R, Gell G, Prokosch H-U (eds.): *Medical Infobahn for Europe: Proceedings of MIE2000 and GMDS2000*, Studies in Health Technology and Informatics, vol. 77, IOS Press, Amsterdam, 2000, pp. 753-757.

V. Laurikkala J: Improving identification of difficult small classes by balancing class distribution. In: Quaglini S, Barahona P, Andreassen S (eds.): *Artificial Intelligence in Medicine: Eight European Conference on Artificial Intelligence in Medicine in Europe*, Lecture Notes in Artificial Intelligence, vol. 2101, Springer, Berlin, 2001, pp. 63-66.

VI. Laurikkala J: An Internet-based multi-expert system for the differential diagnosis of female urinary incontinence, *Medical Informatics & The Internet in Medicine*. (submitted)

Contents

1. Introduction

Human reasoning is inherently fuzzy and somewhat illogical and, unfortunately, physicians are no different from other humans in this respect. Human beings are likely to abandon logic and use heuristics when they are faced with decisions that involve uncertainty [Mac95]. Heuristics are fast rules of thumb that produce correct decisions most of the time, but may also lead to errors [Wat86, Mac95]. The reliability of heuristics under uncertainty depends largely on the skills of the person who applies them. For example, senior physicians tend to perform better than juniors [Nyk00]. Medical decision making is difficult, because it involves humans as the decision makers and other humans as information sources and as the subjects of decisions. Exact sciences, such as computer science, mathematics, or statistics, use formal methods to model, quantify, and control uncertainty. This type of approach is hardly feasible in medicine due to the central role of humans in the profession and its decision making processes.

Physicians' work has many aspects that are likely to cause uncertainty hampering the decision making of human beings. Incomplete data are a common problem in medicine. Patient data may be missing because of haste, by omission, or due to human error. Patients may also have difficulties in describing their condition in words, and, on the other hand, there are patients who are able to describe their symptoms, but their descriptions contradict those they gave earlier. In addition, health professionals are nowadays confronted with an increasing amount of information from different sources. This information overload may affect the quality of patient work, because selection of the relevant information is difficult [Nyk00]. Decision making even in a restricted medical subspecialty involves large amounts of knowledge and information. Cost reduction - another current trend in patient care - forces physicians to work within a tighter schedule, because the same number of patients must be treated with fewer resources.

Computerised decision support systems (DSS) have been used for decades in medicine to aid healthcare professionals [Mil94]. These systems help humans by providing consistent and reproducible reasoning under uncertainty. DSSs have been developed especially to support diagnosis, which is one of the main tasks in medical profession [Mac95, vBM97 Chapter 1, Nyk00]. Even though the opinions on the

effect of DSSs on physician performance on diagnosis vary [Mil94, HHH98+], medical DSSs have become an established component of medical technology [Mil94]. There exist successful applications in areas such as computerised electrocardiogram analysis, cytologic recognition and classification, drug dosing, and preventive care [Mil94, HHH98+]. Moreover, many medical DSSs have been able to perform as well as experienced physicians or even to outperform them [DLS72+, HSB94+, PEJ96, KAJ98+, CFC01+]. This study addresses expert systems, which are advanced DSSs that perform in a narrow domain like a human who has the expert level skills of the area. It is essential that DSSs, such as expert systems, are viewed as decision aids in diagnosis. It is acceptable to use computerised decision methods, but the physician should always be ultimately responsible for the decision [Mil94, vBM97 Chapter 1]. Computers can also be utilised in other ways to help physicians. Computer programs may act as checklists that remind when something essential has not been done [Mil94, HHH98+]. Computer programs are also able to provide feedback to a physician by criticising his or her decisions [GW98]. In addition, computers facilitate physician's gathering of background information from electronic sources [vBM97 Chapter 1].

We studied computer-aided decision support of the differential diagnosis of female urinary incontinence. Urinary incontinence, i.e. involuntary loss of urine, is a fairly common problem in women. The prevalence of urinary incontinence among women between 15 and 64 years of age is 10-25% [Uri92]. Hu [Hu90] estimates that over 10 milliard dollars per year are spent in care for the incontinent patients in the United States alone. The differential diagnosis of female urinary incontinence is problematic for physicians for various reasons. Firstly, diagnosis made on the basis of the patient history alone may be unreliable [JNO94]. Therefore, urodynamic testing is needed for sufficiently accurate diagnosis, especially when symptoms of urgency are present. Secondly, medical knowledge suggests that the incontinence classes are to some degree overlapping. For example, one frequent diagnostic class is mixed incontinence, which has characteristics from both the stress incontinence class and classes where urgency is clearly the dominant symptom. Thirdly, the reliability of the final diagnosis is important to avoid unnecessary surgeries [JNO94].

This study describes an expert system for the differential diagnosis of female urinary incontinence. The system was named IES1 (for Incontinence Expert System, version 1). Existing data mining techniques and our new methods were applied in the construction of the system. We made use of machine learning methods to discover

diagnostic knowledge automatically from the patient cases. The methods included a new learning system based on genetic algorithms. The aim was to avoid the manual knowledge acquisition which is the major difficulty in expert system development. In addition, different data pre-processing methods were applied, and a new one was developed, to enable the knowledge discovery.

We started to study the differential diagnosis of female urinary incontinence, because the diagnostic task is difficult for physicians, and few studies have applied methods of artificial intelligence in this area of medicine. Earlier research includes an expert system developed as a decision support and teaching tool for the diagnosis of female urinary incontinence [RK88] and an interactive expert system that was used to provide information for the incontinent women [Gor95]. In addition, the literature searches made at the beginning of the study indicated no previous research involving data mining techniques in this area of medicine. Data mining in connection with expert systems was interesting from computer scientific viewpoint, because it seemed that the greatest progress in the area of the expert systems might be achieved in the automatic acquisition of the domain knowledge with machine learning methods. A new learning paradigm - genetic algorithms - was appealing, because artificial neural networks (ANN), which are also inspired by nature, had been successfully used in numerous medical applications [FD95, Cha98], such as the diagnosis of acute abdominal pain [PEJ96]. The data pre-processing research was initiated later, when we ran into the practical problems concerning the analysis of real world data.

The introductory part of this study is organised as follows. Section 2 gives the reader background information on knowledge and its acquisition, expert systems, and data mining. The aims of the study are stated in Section 3. Section 4 introduces several issues related to machine learning. First, representation of data and difficulties related to real world data are discussed. Second, genetic algorithms and instance-based learning, which were used in the research involving new methods, are described. Finally, measures for the evaluation of the learning output are presented. The original publications are reviewed in Section 5. The introduction ends with discussion and conclusions in Section 6.

2. Background

This section provides a reader with background information intended to facilitate understanding of the later sections of the study. We first introduce different aspects and definitions of knowledge in Section 2.1. We then discuss the expert systems and their relations to the knowledge-based systems and decision support systems in Section 2.2. Knowledge acquisition and its difficulties are presented in Section 2.3. Lastly, we discuss data mining - the automated process for discovering knowledge.

2.1. Knowledge

Turban [Tur93] classifies data, information, and knowledge by their degree of abstraction and by their quantity. Knowledge is the most abstract and exists in the smallest quantity, while data are the least abstract and exist in the largest quantity. Knowledge can be viewed as information that has been organised and analysed to make it understandable and applicable to problem solving or decision making [Tur93 Chapter 11.4]. Aamodt *et al.* [AM95] characterise knowledge as learned information. Knowledge is an output of the learning process, where information is incorporated in reasoning resources and is made ready for active use within a decision process.

Knowledge has several definitions that characterise it by the extremes of its continuum. Knowledge may be described by its level as deep or shallow knowledge. Deep or scientific knowledge is formal knowledge which deals with deduction - the understanding of general principles and relations. In problem solving humans often utilise deep knowledge which one can draw, for instance, from books and articles or acquire through education. Scientific knowledge is reusable, i.e. it can be applied to solve different problems in different situations. By contrast, shallow or experiential knowledge is related to specific situations or facts. This knowledge one often learns through induction. Consider diagnostic work-up as an example of these knowledge types: In diagnostic work a physician uses knowledge of the biological processes and relations between the pathophysiological conditions and disease symptoms. This deep knowledge allows the physician to diagnose patients with different diseases. On the other hand, the physician may make use of shallow knowledge by recognising a disease on the basis of certain symptoms that he or she has seen before [Tur93, vBM97 Chapter 15, Nyk00].

Knowledge is often categorised into declarative, procedural, and meta-knowledge. Declarative knowledge is descriptive and shallow; it gives facts on the state of the world. Human experts are typically able to verbalise declarative knowledge. Procedural knowledge relates to procedures and sequences that are needed in the problem solving process. It tells how to use declarative knowledge and how to make inferences. Meta-knowledge is knowledge about knowledge. For example, an expert system may have knowledge about its reasoning capability [Tur93, Nyk00]. Another viewpoint on knowledge is to characterise it as tacit or explicit. Skills that one has acquired so well that he or she can no longer explicitly explain them, are viewed as tacit knowledge, while explicit knowledge can be explained in some way [Nyk00].

Representation of knowledge is a central issue in systems that utilise it. IF-THEN rules are the most frequent way to capture knowledge in expert systems. These rules describe the action, the THEN part of a rule, that is performed when the current facts match the conditions, i.e. the IF part of a rule. Frames are another popular way to represent knowledge. The frame-based methods use a network of nodes that are connected by relations and are organised into a hierarchy. Nodes correspond to concepts or objects that have attributes and their values. Frames provide a more natural and flexible way to present knowledge than rules [Wat86].

The researchers of machine learning provide yet another viewpoint on knowledge: They classify knowledge according to its presentation as symbolic and sub-symbolic [MK90]. Machine learning methods produce, for example, symbolic decision rules that a human is able to understand and validate. On the other hand, sub-symbolic knowledge, such as the weights of neural network, is difficult for a human to comprehend. See Section 4.1 for further discussion of this subject.

2.2. Expert systems

Waterman [Wat86] defines expert systems [Wat86, Tur93, Lie98, Nyk00] as computer programs that use expert knowledge to attain high levels of performance in a narrow problem area. The heart of an expert system is a corpus of knowledge, the knowledge base, that allows the system to solve problems as well as the domain expert. The system should produce solutions in the same time as the expert and it should also be able to explain its reasoning. As an institutional memory of knowledge, the expert system may be used to train new personnel. Waterman [Wat86] sees expert systems as a subspecialty of knowledge-based systems (KBS) [Wat86, H-RJ94]. Both

systems store the domain knowledge in the knowledge base, which is separate from the general problem solving knowledge, which is called inference engine. Expert systems, however, use the knowledge as skilled domain experts [Wat86].

Figure 1 shows the structure of a rule-based expert system. The knowledge base contains domain knowledge coded into the IF-THEN rules and facts. The inference engine comprises an interpreter, which decides how to apply the rules, and a scheduler, which gives the application order of the rules. The inference engine uses the rules and the known facts to produce inference chains. The matching of the rules may cause one or more of the rules to fire and, thus, new facts are produced and added to the knowledge base. The successive matching of rules controls the program flow with no need for explicit directions as in conventional programs. Expert systems use heuristic knowledge, that is simplifications that limit the search of solutions. The heuristic method is fast, but may sometimes produce an erroneous solution. The algorithmic method always gives a correct or optimal solution, but requires enumeration of all solution candidates, which is slow and sometimes impossible [Wat86].

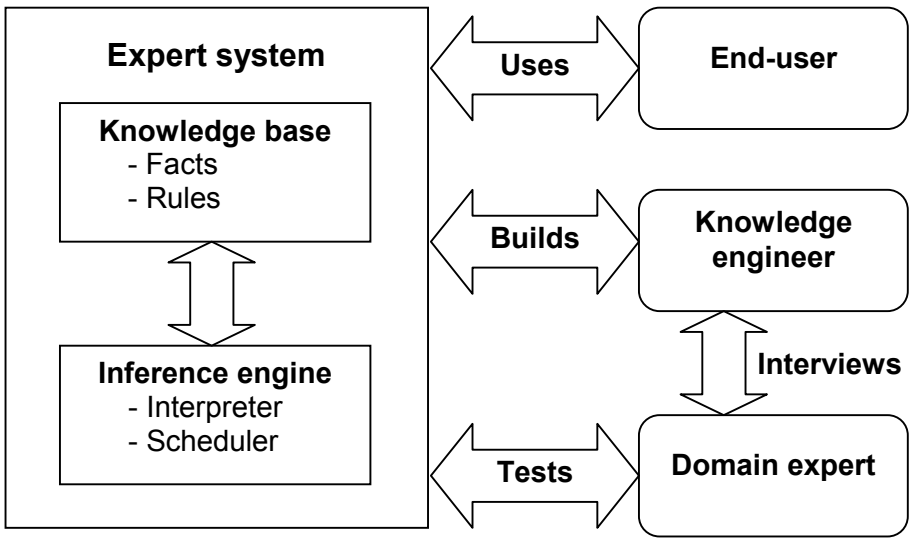

Figure 1. The structure of a rule-based expert system and relations between the system, end-user, knowledge engineer, and domain expert.

The early successes of the expert system technology led the developers to consider the users of expert systems as passive agents who only query answers from the systems. This misinterpretation of the role of the user is known as the Greek Oracle model, which refers to the ancient oracle of Delphi who gave equivocal answers to persons who sought advice [Mil94, vBM97 Chapter 17]. Nowadays, expert systems are considered more like decision support systems [Tur93, vBM97, Nyk00] which help the user in the decision making [Nyk00]. There exists no universally accepted definition for DSS [Tur93]. DSS may, in principle, be any computer program that helps decision makers to make decisions [Tur93, vBM97 Chapter 16, Nyk00]. On the other hand, the characteristics of KBS, such as interactivity and the use of knowledge [Wat86, H-RJ94], are often seen as essential prerequisites for DSS [Tur93, vBM97 Chapter 15, Nyk00]. On the basis of these definitions, we consider in this study an expert system to be a special type of DSS. A notable exception to this relation is given by Turban, who considers expert systems and DSSs to be different disciplines rather than having a nested relation. However, Turban emphasises integration of expert systems and DSSs so that they can complement each other [Tur93 Chapter 1.11].

One of the main application types of expert systems is diagnosis [Wat86, Dur96]. A review by Durkin [Dur96] showed that approximately 30% of expert systems were developed to solve diagnostic problems. When application areas are considered, the dominant areas are business, manufacturing, and medicine [Dur96]. Other large application areas include engineering, power systems, computer systems, and transportation. Since diagnosis is of great importance in medicine, it is not surprising that most of the medical DSSs and expert systems have been developed to aid diagnosis [Wat86, Nyk00]. The early medical expert systems, such as MYCIN [SDA75+, Wat86], were pioneering work which was valuable for expert system research, but few of the systems were ever used in the practice. The systems were difficult to use because of the Greek Oracle approach to decision support and because of the technical limitations.

From the mid-eighties onward, PC technology and expert system shells have allowed efficient building of user-friendly systems. As a result, the expert systems have matured from prototypes to real world commercial applications [Dur96, Lie97], but opinions on the future of expert systems vary. According Durkin [Dur96] and Liebowitz [Lie97] the field is flourishing and new applications are being built increasingly. Conversely, De Hoog [DeH98] argues that very KBS specific

methodologies are in decline and they will be absorbed in the overall system development. The popularity of the medical applications is partly explained by the wealth of well-defined diagnostic problems where the early computerised methods could be applied with relative ease to help physicians [Wat96]. Although expert systems are nowadays built in increasing numbers in other areas, medicine continues to be a popular application area [Dur96, Lie97, Cha98].

2.3. Knowledge acquisition

Knowledge acquisition refers traditionally to a process where a knowledge engineer repeatedly interviews experts to extract knowledge which the experts have gained through their education and practical experience (see Figure 1) [Wat86]. The knowledge engineer codes the knowledge into the knowledge base, for instance, as the IF-THEN rules. In addition, the knowledge engineer may extract knowledge from other sources such as the literature, case studies, and personal experience, but often the expert is the major source of knowledge.

Knowledge acquisition has been identified by various researchers as the major bottleneck in the development of expert systems [Wat86 Chapter 14, Tur93 Chapter 13.3]. Ideally, the expert is a highly skilled person who is open, articulate and motivated, is familiar with computers, and has time for the interviews [Wat86]. Since it is very difficult to find such experts, the knowledge engineer is often faced with a multitude of problems that make knowledge acquisition a laborious, time-consuming, and expensive process. Experts often have difficulties in describing their knowledge in words, they may be busy, or they may be sceptical of the value of using computers and expert systems. Even when an expert is available and articulate, he or she can explain his or her conclusions with plausible lines of reasoning that, however, resemble the actual use of knowledge only remotely. Furthermore, problem solving of the different experts may differ, and, consequently, the extracted knowledge may also differ greatly [Wat86, Tur93].

This problem has been addressed with system-building aids that perform knowledge acquisition semi-automatically or automatically [Wat86, Tur93, Lie98]. Semi-automatic methods support either the expert or the knowledge engineer in knowledge acquisition [Tur93 Chapter 13.5]. Systems such as AQUINAS [Tur93 Chapter 13.10] try to eliminate the knowledge engineer as a mediator in the knowledge acquisition process. These systems facilitate the expert's building of

knowledge bases with no or with little interaction with the knowledge engineer. Systems that aid the knowledge engineer are intended for more efficient execution of common engineering tasks. The automatic methods remove both the experts and the knowledge engineers from the development process or try to minimise their role in knowledge acquisition. For example, data mining where different machine learning techniques are applied allows nearly automatic knowledge acquisition.

2.4. Data mining

Data mining [FP-SS96, HK01, HMS01] has aroused increasing interest in recent years both among researchers and practitioners, as the advances in technology for computing and storage have allowed the collection of large data sets in a well-organised manner. Data sets have been mined, for example, in medicine [ZLK99], biomedicine, finance, marketing, manufacturing, retail industry, and telecommunications [FP-SS96, HK01]. Data mining is often treated as a synonym for knowledge discovery in databases (KDD) [FP-SS96], in industry and media, because the term is shorter and has intuitive appeal [HK01]. On the other hand, data mining can be seen as a major step in the process of knowledge discovery in databases [FP-SS96, HK01]. In this study we shall use data mining as a synonym for KDD because of the convenience and increasing popularity of the term.

Data mining (or KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [FP-SS96]. To put it simply, data mining refers to a process where large amounts of data are mined with computer programs to find ‘golden nuggets’ of knowledge. There are various definitions for the steps of the data mining process (see Figure 2), which often differ in number and sometimes in order [FP-SS96, HK01]. For example, data collection and knowledge utilisation steps are often omitted from the figures describing the process. The core of the process is knowledge discovery with computerised methods, such as machine learning and statistical methods. However, the other steps are as important as the discovery step for the successful application of data mining [FP-SS96]. Preliminary steps include data collection, selection, pre-processing, and transformation. For example, appropriate data are first retrieved from the database, and then possibly cleaned and reduced before knowledge discovery. After the discovery step, the knowledge will be evaluated by humans or on the basis of objective quality criteria. The knowledge may be incorporated, for example, into a

DSS or it may be documented and presented to a wider audience. Data mining is an interactive process which is also iterative, i.e. the whole process or parts of it may be performed repeatedly.

Data mining is frequently used for prediction and description. Prediction involves estimating the values of future examples, while description focuses on finding patterns describing the data [FP-SS96]. Data mining has connections with conventional statistical analysis, but there are also fundamental differences [HMS01]. First, data are usually mined from data sets that are considerably larger than data sets in the statistical analyses. Also, data mining is considered as secondary data analysis utilising data that have usually been collected for some other purpose. Statisticians perform primary analysis, where data are carefully collected to test hypotheses. Last, data miners work with real world data that have missing values, noise, outliers, and other difficulties.

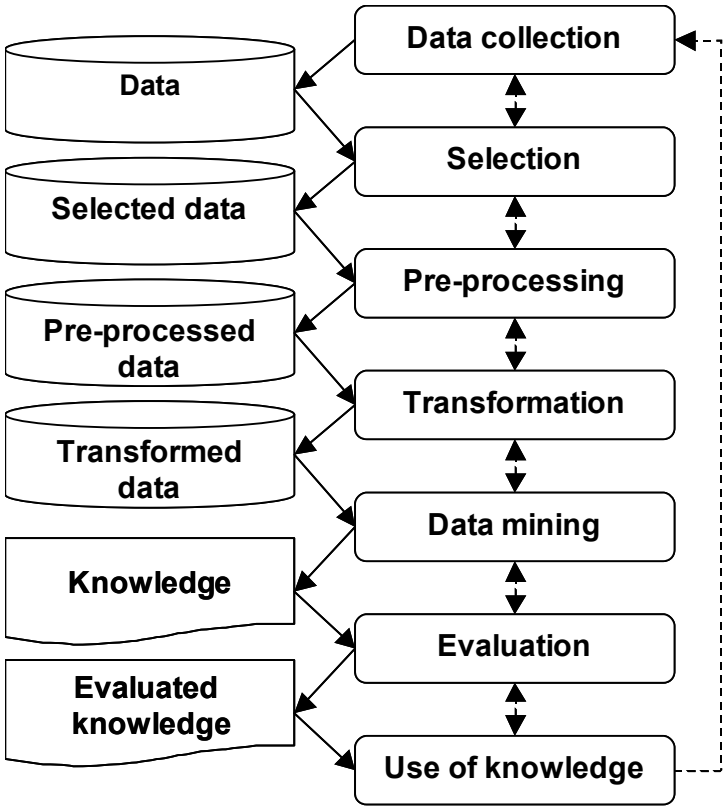

Figure 2. The data mining process. Solid lines show the input data to and the output data from the different steps of the process. Dashed lines indicate the passing of control from one step to another.

3. Aims of the study

The aims of the present study were both practical and methodological. The practical aim was to construct an expert system as a decision support tool for physicians in the differential diagnosis of female urinary incontinence. The methodological aim was to study whether the expert system could be developed using data mining methods to overcome the knowledge acquisition bottleneck (see Section 2.3). We designed new data mining techniques, inspired by the practical problems, and tested them especially in the context of mining female urinary incontinence data. To achieve the practical aim, we applied existing machine learning and data pre-processing methods, and developed new ones, to enable the mining of diagnostic knowledge for the expert system.

To summarise, the main aims of this study were:

- to develop an expert system for the differential diagnosis of female urinary incontinence
- to research new data mining methodology for solving real world problems.

The main aims can be decomposed into the following sub-aims:

- treatment of missing data
- analysis of the usefulness of attributes
- research of genetic algorithms in machine learning
- identification of noise and outliers
- balancing of the imbalanced class distribution
- use of data mining to discover diagnostic knowledge automatically for the expert system
- implementation of the expert system as an Internet-based application.

The present study consists of the introduction and of six original publications (Appendices I-VI). The author was the main contributor and responsible author in the original papers (I)-(IV), which were prepared in collaboration with the co-authors. The co-authors' contribution to the papers was the following. Jorma Penttinen and

Pauliina Aukee defined the diagnostic parameters, described the diagnostic work-up, and evaluated the results with the author in paper (I). Seppo Lammi helped the author in statistical issues in papers (I) and (III). Kati Viikki kindly introduced the decision trees to the author in paper (III). Martti Juhola commented all the manuscripts and helped the author in methodological issues related to computer science.

The original work includes five papers that are reprinted in their published format with the kind permissions of the publishers. One paper (VI) is a submitted manuscript. The papers are presented in their preparation order. Paper (I) was our first work, but it was published in 2001 because of a long review process. The chronological order of the other papers corresponds to their preparation order. Paper (I) presents a statistical evaluation of the attributes of the female urinary incontinence data set where missing data were filled in with different methods. Paper (II) describes a learning method that makes use of genetic algorithms to produce classification rules. In paper (III), the classification capability of the genetic algorithms is compared with different machine learning and statistical methods. Paper (IV) involves work where different proximity functions were compared to identify the best one for the research of instance-based methods. An instance-based method for balancing of imbalanced class distribution is presented in paper (V). The last paper (VI) describes an Internet-based multi-expert system for the differential diagnosis of female urinary incontinence.

4. Machine learning

In this section, we first give an introduction to the field of machine learning and different categories of learning methods, and briefly describe the C4.5 decision tree generator and ANNs. Second, a data matrix consisting of examples and attributes is presented. Third, real world data related problems that we encountered during the research are discussed. We present then in greater detail genetic algorithms and nearest neighbour classifier - the learning methods that were used in our studies of new data mining techniques. Lastly, different measures for evaluation of the classification ability of the machine learning methods are discussed.

4.1. Introduction

Machine learning [MK90, Mit97] is a sub-domain of artificial intelligence concerned with developing computational theories of learning. Broadly defined, machine learning includes any computer program that improves its performance at some task through experience [Mit97]. Machine learning has been applied in various areas including data mining, speech recognition, computer vision, robotics, and game playing [MK90, Mit97]. Learning to drive an autonomous vehicle, classification of new celestial objects, playing of world-class backgammon, making credit decisions, diagnosis of mechanical devices, reducing banding in rotogravure printing, preventing breakdowns in electrical transformers, and forecasting of severe thunderstorms are only some examples of specific applications [LS95, Mit97]. Machine learning has been used extensively in the area of medicine: Some of the medical applications are described in [FNI91, Con95, FD95, PEJ96, Cha98, II, Tsu98, KLP99+, Lav99, ZLK99, GE00, CFC01+, VI]. Machine learning is inherently a multidisciplinary field that has been influenced by areas including artificial intelligence, probability and statistics, computational complexity theory, information theory, philosophy, and psychology [Mit97]. The machine learning community is increasingly interested in statistics, because many well-known statistical methods are often applicable in the area of machine learning. Statistics is needed in the evaluation of hypotheses and results, and in comparison of methods [GMP96+, Mit97, Sal99].

Machine learning methods may be defined according to their primary purpose as synthetic and analytic [MK90]. Synthetic learning aims to produce new or better knowledge, while analytic learning reformulates knowledge into a better form.

Michalski *et al.* [MK90] use as the further classification criteria type of learning input, type of primary inference, and the role of prior knowledge. The majority of learning methods may be described as synthetic methods that learn from examples by induction, and are empirical, i.e. use little background knowledge in learning. We shall refer to this learning type as empirical induction from examples (EIE). Many rule and decision tree induction programs, supervised neural networks, and genetic algorithm based learners fall into the EIE category, where inductive inference is understood narrowly as empirical generalisation of examples without using much prior knowledge. On a more abstract level, induction is the opposite of deduction. While deduction is derivation of consequents from given premises, induction hypothesises premises that entail given consequents [MK90]. Induction systems utilise inductive bias - prior assumptions regarding the task - to generate knowledge [Mic97 Chapter 2.7]. For example, an approximate inductive bias for a decision tree learner might state that smaller trees are better than larger trees [Mic97 p. 63].

As the multivariate statistical methods, machine learning may be used both for the descriptive and predictive analysis of data. In descriptive analysis the aim is to produce knowledge, i.e. models, that help to better understand the underlying regularities of data. Predictive analysis aims for the accurate identification of new unseen examples. Machine learning is often applied to solve classification problems, where the task is to assign each example to a class. It is usually assumed that the classes are mutually exclusive and the number of class labels is limited [Qui86]. Knowledge used for classification is often referred to as a classifier, which takes as an input the attribute values and gives as an output a classification [Qui93]. Categorisation of machine learning methods to supervised and unsupervised learners originates in the availability of class information during learning. Most machine learning methods are supervised methods which use the class information during learning. However, unsupervised methods such as cluster analysis [JD88, Eve93] work only with the data and do not use the class information.

Learning methods may also be categorised according to the knowledge presentation as methods that produce symbolic and sub-symbolic knowledge [MK90]. Symbolic knowledge is easier for humans to understand than knowledge represented on sub-symbolic level. Symbolic knowledge often conforms to the comprehensibility principle [MK90] according to which the knowledge created by the learning program should be in a form that is easy for a human to interpret and comprehend.

Comprehensible knowledge presentation employs a limited number of terms and operators that correspond to those that human experts use [MK90]. For example, decision rules and trees are symbolic knowledge, while neural networks are often mentioned as examples of sub-symbolic systems. Symbolic representation is not necessarily always better than sub-symbolic representation. The learning task may be such that the symbolic representation is too cumbersome and complex for humans [Qui93 p. 45], and, sometimes, it is more important to learn a good solution than to explain it [LS95]. Symbolic learning systems may also categorise quantitative attributes in a different way than the experts are familiar with [Tur93 p. 535]. Nevertheless, there exist problem domains, such as medicine, where people need to understand the system behaviour [MK90, FD95, Lav99]. Furthermore, real world applications of machine learning have shown that knowledge which is comprehensible for a user is an important factor in the development of successful applications [SN98].

A detailed description of the various learning methods is beyond the scope of this work. An interested reader will find a good introduction to some currently used methods in [Mit97]. Michalski *et al.* [MK90] discuss the current and early work in the area of machine learning. A number of machine learning methods is briefly presented and compared for their accuracy, complexity, and training time in [LLS00]. In the following, only C4.5 decision tree generator and ANNs, which nowadays are probably the most popular learning methods, are briefly introduced. Genetic algorithms and nearest neighbour classification are discussed in Sections 4.4 and 4.5.

C4.5 [Qui93], the descendant of ID3 [Qui86], is a widely used decision tree generator that was also applied in this work. C4.5 takes as its learning input examples described with attributes having a categorical or quantitative domain. The learning output is a decision tree where leaves represent classes and nodes are tests based on attributes. The decision tree is constructed using a top-down approach starting from the root of the tree and applied recursively until the tree is complete. At each step of the building process, the attribute which divides the training set in the best possible way in terms of gain ratio is selected to be the test of the node. Overly complex decision trees can be reduced by pruning, and when further simplification is needed, unpruned trees may be converted into rules. The conversion into rules is not straightforward knowledge reformulation, because rules are also generalised by deleting conditions which seem to be irrelevant to the classification [Qui93].

ANNs [RWL94, Swi96] are a biologically inspired learning method that is frequently used to solve difficult tasks that involve uncertainty. A neural network is a cognitive model capable of learning and composed of processing elements (nodes of network) and connections. Weights, which are distributed among the connections, determine the propagation of excitatory and inhibitory signals that define the excitation of the nodes. The most common method for adjusting the weights is the backward propagation algorithm [RWL94, Swi96]. The knowledge that a neural network has learnt is represented as a sub-symbolic form by the weights the network. Even though the sub-symbolic knowledge may be partially explained [Swi96 Chapter 7], the neural networks are often black boxes whose functioning is difficult or impossible for a human to understand [HW90, FD95].

4.2. Representation of the data

The data are represented conventionally in statistics as a data matrix. Many machine learning methods, such as the methods used in this study, also assume that the data are in a matrix form. This representation organises the data as an $n \times m$ matrix where the n rows represent examples (cases, instances or objects) and the m columns correspond to attributes (variables or features) [Qui86, Mit97, HMS01]. In data mining, the data may differ from this classical representation; data may, for example, be text or images [HK01, HMS01]. Each example has a class label, i.e. the data matrix also has a class attribute that gives the class of an example. Usually, the class of interest is referred to as the positive class and the other classes are known as the negative classes. It should be noted that only the data attributes are presented to unsupervised learners.

The scale of the attributes indicates how much information the attribute contains. Attributes with nominal and ordinal scales are categorical [Agr96]. Nominal attributes have no order for their values, while ordinal attributes have order, but the successive categories do not represent equal differences in the scale. Both nominal and ordinal attributes or only nominal attributes are called qualitative attributes depending on the source. In this study, we use qualitative attributes as a synonym for categorical attributes and state explicitly when a narrower definition is used. Quantitative attributes are described with interval and ratio scales. The values of the interval scaled attributes have order, and differences of successive values are equal. Ratio scale has all the properties of the interval scale, and, in addition, it has a natural base value that cannot be changed [Agr96, Sha96, HMS01].

As an extremely simple example, consider a problem where we must classify healthy persons and persons suffering from influenza into 'healthy' and 'sick' classes of which 'sick' is the positive class. A physician has diagnosed the patients on the basis of data he or she has gathered by examining the patients. Suppose that the physician has recorded the following data for each patient: sore throat (values {'no', 'yes'}), headache (values {'no', 'slight', 'moderate', 'severe'}), temperature, sex (values {'female', 'male'}), and age. The diagnosis is the class attribute, and sore throat, headache, temperature, sex, and age are the data attributes. Table 1 shows the data matrix.

Table 1. A data matrix with a class attribute (diagnosis) and five data attributes.

Diagnosis	Sore throat	Headache	Temperature	Sex	Age
healthy	no	no	36.5	female	45
sick	yes	severe	40.2	male	65
:	:	:	:	:	:
healthy	no	slight	37.1	male	30

Sore throat and sex are nominal attributes, because there is no meaningful order for their values. Headache attribute is categorical, but since there is a natural order for the degree of the headache, the scale of this attribute is ordinal. Temperature (C^o) is a classic example of the interval scaled attribute. Age is measured with ratio scale, because it has a natural base value (0). Since the base value of temperature is arbitrary (freezing point of water), its scale cannot be a ratio scale. The previous example showed that some attributes may have less value in learning: The sex and age of a patient are probably of little use when the task is to induce rules for the classification of patients without or with influenza. Irrelevant attributes and other problems that may hinder machine learning are discussed in Section 4.3.

4.3. Difficulties with real world data

There are various real world data related problems in applying machine learning and many of them are the same as in statistics or very similar to the well-known statistical problems. However, machine learning methods are usually less sensitive to difficulties caused by real world data than statistical techniques. Problems that we encountered during the research were mixed attributes, irrelevant attributes, missing

values, unusual data, and imbalanced class distribution. In the following, we discuss these problems from the machine learning and statistical viewpoints. Saitta *et al.* [SN98] discuss the more general problems that one is likely to encounter in projects that use machine learning methods in real world.

4.3.1. Mixed attributes

Many statistical methods assume that the data have certain scale and distribution. For example, linear regression and discriminant analysis expect quantitative attributes with multivariate normal distribution [Wei85, Sha96]. Likewise, the machine learning systems make assumptions of the scales of attributes. For example, ID3 requires categorical data [Qui86], while the nearest neighbour method, with Euclidean distance, assumes that the data are quantitative [Mit97]. Unfortunately, real world data do not often meet the given requirements. Statistical as well as machine learning methods are especially difficult to apply when the data are mixed, i.e. described by both categorical and quantitative attributes.

One approach to overcome this problem is to reduce the quantitative attributes into categorical ones. However, reduction may result in serious loss of information, and, moreover, the objective definition of the categories may be difficult. Sometimes it is possible to make use of the order information in ordinal attributes by converting them into quantitative attributes by assigning numerical scores to the categories [Agr96]. This approach is not applicable for nominal attributes, because it is impossible to establish order for these values. The values of nominal attributes can be coded into new binary attributes (dummy variables) which some methods assuming quantitative data can process [Wei85, Agr96]. The drawback of using dummy variables is the increase in the problem size: As the number of attributes increases, additional data may be needed to keep the classes large enough for the analysis method.

4.3.2. Irrelevant attributes

The selection of the best subset of the available attributes for the analysis of data is an important area of research both in statistics and machine learning. Selection of relevant attributes is a central issue in multivariate statistical analyses where simpler models are preferred over the more complex ones [Wei85, Sha96]. Removing of irrelevant attributes is often needed in machine learning to reduce computation time, because machine learning is typically applied to large data sets. In addition, a reduced

subset of features allows a learning algorithm to decrease the number of hypotheses under consideration (to reduce search space), and, thus, to produce more general concepts [BL97, DL97]. Some methods, such as stepwise multivariate statistical methods and C4.5, are able to reject the irrelevant features. On the other hand, there exist methods that use all the attributes throughout the analysis. The simple nearest neighbour method and many clustering techniques do not modify the attribute set [Eve93, Mit97]. Since automatic attribute selection during analysis has its dangers, for instance exclusion of good predictors due to multicollinearity [Agr96, Sha96], one can try to identify the relevant attributes before the actual analysis. Discussions with the experts may give valuable information on the usefulness of the attributes [LS95], and simple statistical evaluations of dependencies are often of great help. There are also available a number of computerised methods, such as Relief and its extensions, for the prior selection of attributes [BL97, DL97].

4.3.3. Missing values

The real word data frequently have missing values that have numerous origins. Known values may be missing because of omission or haste, or because they were considered unimportant. Missing values are often missing simply because they were never acquired. For example, some of the diagnostic tests may not be needed to reach the final diagnosis. In addition, the method used to collect data affects the amount of missing data. In prospective studies the data are usually collected systematically, while retrospective data collection results in some missing data, since it is usually impossible to acquire the missing values later. Unlike statistical methods, many of the machine learning methods, such as ID3, C4.5, ASSISTANT [Lav99], and some rule induction systems including [DJSG93, Jan93, II], have a built-in capability to treat missing values. However, there exist learning methods that require complete data. For example, neural networks assume that the input data have no missing values.

The main approaches to treat missing data before the analysis are complete-case analysis, available-case analysis, and imputation [LR87, SO98]. In complete-case analysis only complete cases are used. This approach is feasible with few missing values, but otherwise a large amount of the data may be lost, and the consequent analyses may be biased [LR87]. In available-case analysis, cases that have values for the attributes involved in the analysis are used. This approach utilises more data than the complete-case analysis, but the analysis of the results may be difficult due to the

different number of cases in different analyses. Lastly, the missing values may be imputed (filled in) without biasing the data when the missing value mechanism, i.e. the process that produced the missing values, is ignorable for the imputation method [LR87, SO98].

4.3.4. Unusual data

Another frequent problem in data sets are unusual data: outliers and noise. Outliers [BC83, BL87] are observations which appear to be inconsistent with the remainder of the data. Human error often produces unintentional outliers. Outliers are also frequently generated as a result of the natural variation of population or process one cannot control [BL87]. Univariate outliers are extreme data values of distribution of an individual attribute, while multivariate outliers are examples which have unusual value combinations. Multivariate outliers are not necessarily outliers in a univariate sense, because combinations of normal values may be abnormal. Noise is mislabelled examples (class noise) or errors in the values of attributes (attribute noise) [Qui86 p. 92]. Outlier is a broader concept than noise, because it includes errors as well as discordant data produced by the natural variation of population. Examples with class noise are outliers produced by sampling error, while attribute noise may or may not show in the data as outlying values. Outliers and noise pose a problem to all machine learning and statistical methods. The statistical community has studied outliers, because many statistical methods, such as linear regression analysis, are sensitive to outliers [Wei85]. Machine learning methods are able to withstand noise and outliers to a varying degree. Decision trees are quite robust, but noise may cause the attributes to become inadequate and may lead to unnecessarily complex tree structures [Qui86].

4.3.5. Imbalanced class distribution

When some classes are heavily under-represented, many classification methods are likely to run into problems [KM97]. Examples of the small classes are lost among examples of the more frequent classes during learning, and, consequently, classifiers such as decision rules or trees are unable to correctly classify new unseen cases from the minority classes. Moreover, imbalanced class distribution may hamper descriptive analysis, where models describing the data are constructed. The models may give an inadequate picture of the data, if the knowledge from the small classes is not fully included into them. The learning task is even more problematic if the small class is

difficult to identify not only because of its size, but also because of its other characteristics. One approach to overcome imbalanced class distribution is the reduction of large classes before the actual analysis [KM97]. Other approaches include generating artificial data [Swi96, KM97], weighing training cases [KM97, KHM98], and introducing different misclassification costs [KM97, KHM98]. There also exist methods that are insensitive to the underlying class distribution in the training set [KHM98].

4.4. Genetic algorithms

Genetic algorithms [Gol89, Dav91a, Gol94, Mit96, Mic96, BFM97, Mit97 Chapter 9] are robust search algorithms that are loosely based on the principles of natural selection and natural genetics. Genetic algorithms are robust, i.e., capable of good performance in a variety of environments [Gol89]. Genetic algorithms can easily be introduced into new problem domains and existing systems, because their operation requires only a very small amount of problem-specific knowledge [Gol89, Gol94]. A drawback of domain independence is that a genetic algorithm sometimes achieves only a near-optimal performance level, but this problem can be tackled by exploiting problem knowledge [Gol94]. A more efficient search is achieved, for example, with problem-specific operators and codings or by hybridising a local search method with a genetic algorithm [Gol89, Dav91a]. Genetic algorithms have been applied to a considerable number of difficult problems in different areas [Gol89, Dav91a, Gol94, Mit96, BFM97] which include optimisation, machine learning, economics, ecology, evolution and learning, and social systems.

The four classic properties [Gol89] of the genetic algorithms are the sub-symbolic coding of the problem, search from the population of chromosomes which are also known as solutions or individuals, ‘blind’ search based only on the fitness of the chromosomes, and use of stochastic operators. Figure 3 illustrates the functioning of a simple genetic algorithm [Gol89] which is a basic form of the genetic algorithms. There exist a number of more advanced algorithms that differ more or less from this simple algorithm. First, the initial population, which usually consists of binary strings, is created. Then, the fitness of each chromosome in the population is calculated. The fittest chromosomes are selected as the parents for the next generation. Pairs of children are produced by exchanging the parts of parents with the crossover operator.

In addition, small changes are made to the children with the mutation operator. The genetic algorithm loops until the pre-determined terminating condition is fulfilled.

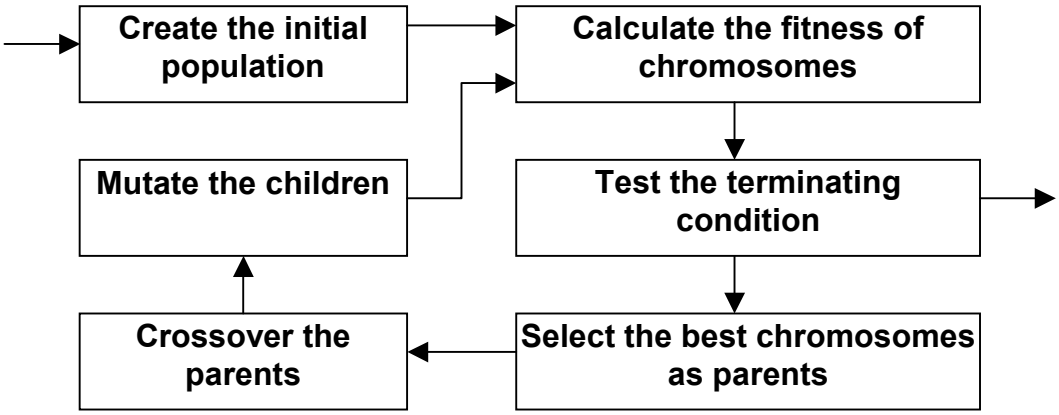

Figure 3. The simple genetic algorithm.

Machine learning systems that utilise genetic algorithms, or genetics-based machine learning (GBML) [Gol89, Mic96, Mit97 Chapter 9], have been researched since the 1980s. In comparison with neural networks and decision trees, GBML is quite a small area within machine learning. However, GBML systems have been used successfully in various areas including the learning difficult multiplexer problems, learning of protein structures, and identification of individuals at risk for coronary artery disease [Gol89, Con95, Mit96]. GBML systems are commonly classified according to the solution representation to the Michigan and Pittsburgh approaches. In the Michigan approach the solution is represented as a set of fixed-length chromosomes. These systems, known as classifier systems, were originally developed by Holland *et al.* [Gol89, Jan93]. In the Pittsburgh approach an individual member of a population consisting of a set of variable-length chromosomes is a solution to a problem. Pittsburgh type systems, also known as learning systems, were proposed by Smith [Gol89, Jan93].

GBML systems can also be grouped according to the problem knowledge that has been included in the system. Some researchers rely on classic domain independent design, while others prefer to include domain knowledge in their systems. An example of the minimal approach is the simplest version of the GABIL system [DJSG93] which uses only the mutation and crossover operators. On the other hand,

the GIL system [Jan93] represents the completely opposite approach with many task specific learning operators. There also exist GBML systems that use fuzzy logic to handle cognitive uncertainties, such as vagueness and ambiguity, involved in classification problems [YZ96].

4.5. Nearest neighbour method

The nearest neighbour method [Mit97] is the most basic of the instance-based learning methods [AKA91, Mit97], which simply store training examples and postpone generalisation until a new instance (or example) is classified. Due to the delayed generalisation, the instance-based learning methods are also known as 'lazy' learners. These methods do not produce general and explicit knowledge. The knowledge that the nearest neighbour method generates is a function that maps examples into classes [AKA91]. The inductive bias of the nearest neighbour method is the assumption that the classification of an example is similar to the classification of nearby examples [Mit97].

The nearest neighbour method uses measures of similarity or distance to determine the classification of a new example. These measures are known in the area of cluster analysis as proximity measures [Eve93]. The most common proximity measure is the Euclidean distance, which assumes a real-valued space [Eve93, Mit97]. There are various other proximity functions designed for different types of data such as Jaccard's coefficient for binary data [Eve93, WM97]. *K*-nearest ($k = 1$) neighbour method (*k*-NN) classifies a new example into the class of the example that is nearest or the most similar to the new example. When $k > 1$, k is usually a small odd integer, for example 3, 5 or 7, and majority voting is often used to select the class. Ties may be broken randomly, for example. Using of $k > 1$ makes the method less sensitive to noise points [Mit97].

The proximity measure should be appropriate for the data. Metricity of the proximity measure, normalisation of the data, and treatment of missing values are important issues in applying the nearest neighbour classification. The framework presented in [MK90] does not include the instance-based learning methods. However, EIE category seems to be the most appropriate for the simple nearest neighbour method, because EIE is the most straightforward type of learning in the framework.

4.6. Evaluation of the learning output

Learned knowledge may be evaluated by subjective or objective criteria, or by both. The knowledge is often presented to the domain expert for subjective evaluation and validation [LS95, SN98]. A human may comment the usefulness and novelty of the knowledge, and the degree to which it corresponds to his or her own knowledge. For example, a physician may inspect a decision tree to find out whether the expected symptoms and tests are included in the tree. Objective evaluation is often related to the estimation of the quality of a solution to a classification problem. The most common measures for classification in the area of machine learning are accuracy and error rate [Lav99]. True positive rate (TPR), true negative rate (TNR), and receiver operating characteristic (ROC) curve are other common measures for performance evaluation. Accuracy and error rate measure the overall classification ability in different perspectives. While accuracy indicates the percentage of correctly classified examples, error rate gives the percentage of misclassified examples. To avoid redundancy, only accuracy is used here.

True positive rate is the ratio of true positive (correctly classified positive examples) to all the positive examples:

$$TPR = N_{TP} / N_P \cdot 100\%,$$

where N_{TP} is the number of true positive examples and N_P is the number of positive examples. Similarly, true negative rate is the ratio of true negative (correctly classified negative examples) to all the negative examples:

$$TNR = N_{TN} / N_N \cdot 100\%,$$

where N_{TN} and N_N are the numbers of true negative examples and negative examples, respectively.

Accuracy is defined as the ratio of correctly classified examples to all the examples:

$$accuracy = (N_{TP} + N_{TN}) / N \cdot 100\%,$$

where N denotes the total number of examples.

These measures may be calculated from the whole data, but usually a part of the data is set aside for testing and the rest of the data is used for training of the learning algorithm. Division of data allows more realistic estimates of the classification ability, because the learning algorithm may overfit the training data, i.e. the classification knowledge is too specific [RWL94, Mit97 Chapter 3.7], and, consequently, classification measures give overly optimistic results. Cross-validation has become a standard method for evaluating knowledge. The process of *k*-fold cross-validation involves splitting of data into *k* disjoint subsets and using each of the subsets as a test set and the union of other *k*-1 subsets as the training set [Qui93 pp. 89-90]. An older approach is to split the data into separate training and testing sets, for example, in 70:30% or 60:40% ratio. Nowadays, more testing sets, for example $k = 10$, are usually used in the cross-validation process. In this study, we refer to accuracies in the training and test sets as descriptive and prediction accuracies respectively.

The use of accuracy has been criticised in the evaluation of real world applications [KM97, KHM98] and in the comparison of learning algorithms [PFK98]. The practical problem of accuracy is that it does not indicate how the positive examples were classified. This is a serious deficiency especially when the class distribution is imbalanced so that the positive class is small compared to the negative classes [KM97, KHM98]. Even though all of the positive cases are misclassified, the accuracy may be almost 100%, because the majority of the negative cases are classified correctly. TPR and TNR indicate how the positive and negative classes were identified. Van Bemmel *et al.* [vBM97 Chapter 15] state that medical decision support models should be analysed at least using TPR and TNR or their complements. TPR and TNR are known in the area of medicine as sensitivity and specificity respectively [vBM97 Chapter 15]. Both van Bemmel *et al.* [vBM97 Chapter 15] and Provost *et al.* [PFK98] prefer the use of ROC curve to single measures of performance, because the curve allows evaluation of classification performance with different decision thresholds. We did not use ROC curves, because ROC analysis in multi-class problems is difficult and because the interpretation of ROC curves is not always a straightforward task [PFK98].

5. Results

The practical aim of our study was to construct an expert system as a decision support tool in the differential diagnosis of female urinary incontinence. Studies reported in papers (I)-(V) ultimately served this aim. When viewed from the context of the data mining process, our work involves mainly data pre-processing, data mining with machine learning methods, and knowledge utilisation. In Section 5.1, the data collection described in [I] is reviewed. Section 5.2 reviews our data pre-processing work: treatment of missing values [I], evaluation of the diagnostic parameters [I, VI], identifying the appropriate proximity function [IV], and a method for balancing the imbalanced class distribution [V]. The Galactica system [II], which uses genetic algorithms in learning, is discussed in Section 5.3. A comparison of Galactica with different machine learning and statistical methods [III, VI] is reported in Section 5.4. Lastly, expert system IES1 [VI] is described in Section 5.5.

5.1. Data collection

We started the work with data collection, because there was no centralised electronic data storage for the data which we were interested in. Firstly, the diagnostic parameters (or attributes) were defined and a diagnostic classification, which consists of the most common female urinary incontinence diagnoses, was constructed. Secondly, a retrospective investigation was performed on the women who were treated because of urinary incontinence.

Sixteen attributes that are relevant in the differential diagnosis of the female urinary incontinence were identified in paper (I) on the basis of the interviews with the experts. This attribute set A_0 consists of urine in the vagina (UVA), urgency score (US), post voiding residual (PVR), probability of motor urge incontinence (PMU), cystometry (CYM), pressure transmission ratio (PTR), minimum urethral closure pressure (MUCP), stress sign (SS), mobility of urethrovesical junction (UVJ), uroflowmetry (UF), cystoscopy (CYP), stress symptom (SSY), continuous loss of urine (CLU), difficulties with voiding (DV), urge symptom (USY), and age attributes. The attributes were mixed, i.e. there were both binary and quantitative parameters. The classification of patients included ‘normal’ class for continent patients and six female urinary incontinence diagnoses: stress, mixed, sensory urge, motor urge and overflow incontinence, and fistula. After defining the classification, the expected

value for each attribute within diagnostic groups was determined by the expert physicians.

We collected retrospectively D_0 (N = 530) and D_1 (N = 65) female urinary incontinence data sets from patient records in the Department of Obstetrics and Gynaecology of Kuopio University Hospital, Finland. Stress urinary incontinence (N = 323 + 49 = 372) was the most common diagnosis in both the data sets. The frequencies of mixed (N = 140 + 10 = 150), sensory (N = 33 + 2 = 35) and motor (N = 16 + 3 = 19) urge incontinence and normal, i.e. continent, patients (N = 18 + 1 = 19) were considerably lower. The only fistula case was excluded from D_0.

A motor urge case that was included in D_0 in papers (II) and (III), was excluded from D_0 in papers (I) and (IV)-(VI), because the patient was young in comparison with the rest of the patients. This inconsistency could not be avoided, because paper (I) was published after papers (II) and (III). However, inclusion or exclusion of a single patient did not affect the results. Since the data set D_1 was collected later than D_0, it was used only in papers (I) and (VI). In this study, for the sake of clarity, we shall use frequencies and descriptive statistics of data set D_0' with 529 patients. The ages of the women in D_0' were 26-89 years, with a mean age of 52.3 years and a standard deviation of ±11.3 years. In D_1 the ages were 34-83 years, with a mean age of 54.3 years and a standard deviation of ±10.9 years.

5.2. Data pre-processing

An informal preliminary analysis with descriptive statistics, frequencies, and contingency tables showed that the collected data needed pre-processing to enable statistical analyses, knowledge discovery, and building of the expert system. We found that attributes for identifying the rare diseases were irrelevant because of the absence of these cases. It was also obvious that we had to evaluate the usefulness of the other attributes and to study whether the attributes were related to diagnoses as expected. Statistical multivariate methods seemed to be the best choice for the analysis of the attributes. However, missing values had to be treated first, because the statistical methods that were the most suitable for the analysis could not process missing data. Another obvious problem was the mixed scales of attributes. Part of the attributes was binary attributes, and the rest were quantitative ones.

The initial classification experiments showed that the imbalanced class distribution and unusual data (noise and outliers) might be problematic. The experiments

confirmed medical experts' knowledge of the overlapping class borders. These difficulties are demonstrated in Table 2 which shows one-nearest neighbour (1-NN) classification results for the imputed data set D_0'. The sensory urge class was the most difficult to identify, because a large number of the sensory urge incontinence cases were misclassified into the mixed incontinence class. In addition, many stress incontinence cases were assigned to the mixed incontinence class and vice versa. Normal patients could easily be distinguished from the incontinent ones. Descriptive accuracy and TPRs for the stress, mixed, sensory urge, motor urge, and normal classes were 85%, 92%, 78%, 58%, 67%, and 100% respectively.

Table 2. One-nearest neighbour classification of the data set D_0'. The three cells with the highest misclassification frequencies are shown in the bold font.

	Predicted class					
True class	Stress	Mixed	Sen. urge	Mot. urge	Normal	Sum
Stress	296	**24**	3	0	0	323
Mixed	**22**	109	7	2	0	140
Sensory urge	2	**12**	19	0	0	33
Motor urge	0	3	2	10	0	15
Normal	0	0	0	0	18	18
Sum	320	148	31	12	18	529

5.2.1. Treatment of the missing values

UF and CYP had the highest missing value rates. The missing data rates in D_0' and D_1 data sets were 97.0% and 90.8% for UF and 90.4% and 86.2% for CYP respectively. Missing data were inevitable for these measurements, because they are needed to confirm the rare overflow and fistula incontinence diagnoses [I]. Since none of the patients had the rare diagnoses, UF and CYP were excluded from the imputation, multivariate statistical analyses, and data mining in all the papers. Exclusion reduced the percentage of missing data in D_0' from 27.4% to 17.9% and in D_1 from 17.5% to 7.4% respectively. After exclusion PMU (63.3%), UVJ (36.1%), and PTR (34.6%) had the most missing values in D_0', while PMU (41.5%), PTR (21.5%), and MUCP (20%) had the highest missing value rates in D_1. The number of complete cases in D_0' increased from 0 to 87 [I].

In paper (I), we applied means, regression, and expectation-maximization (EM) imputation methods to fill in missing values in D_0'. In addition, complete-cases analysis was performed. The results showed that although the imputed values had a moderate agreement, the multivariate analysis produced similar results. Complete-case analysis gave clearly insufficient results. Moreover, the cross-validation of the logistic regression equations in D_1, where the expert physicians replaced missing values, showed that it is possible to classify unseen patients accurately with statistical models obtained from data sets imputed with the three different methods. In papers (II) and (III), rounded means were used to impute missing values instead of the EM method. This approach was reasonable on the basis of paper (I), but the EM method could not be applied in the papers (II) and (III), because paper (I) was at that point still in the review process. The EM imputed data have been used in the latest papers, (V) and (VI), because the EM method biases the female urinary incontinence data, which is assumed to be missing at random (MAR) [I], less than sampling-based methods [LR87]. Data have not been imputed in all the papers, because the machine learning methods used in this study are able to use data with missing values.

5.2.2. Evaluation of the diagnostic parameters

The purpose of paper (I) was to statistically evaluate the parameters, i.e. symptoms, tests and measurements, which were needed to implement the expert system. The missing data values were imputed (see Section 5.2.1). The data were analysed with complementary statistical techniques. Logistic regression analysis was performed to reveal the relations between the diagnoses and the parameters. Hierarchical cluster analysis with six different techniques was conducted to study whether the patients could be grouped with parameters alone into the clusters corresponding to the diagnostic classes.

Logistic regression analysis showed that the set A_1 = { US, PMU, MUCP, SS, UVJ, SSY, USY, age } of attributes ($A_1 \subset A_0$) had significant ($p < 0.05$) relations to the stress, mixed, and sensory urge diagnoses. The motor urge and normal classes were too small for the statistical analysis. Attributes were mostly those that are important in the diagnostic work-up and, moreover, the directions of relations were as expected. The results of paper (I) as well as those of the latter papers showed that the attributes are sufficient for accurate supervised classification of the data. None of the parameters were dropped from the expert system, because the rare fistula and

overflow cases should also be detected. However, in the data mining we used set A_2 = $A_1 \cup \{$ PVR, PTR, CYM $\}$ ($A_1 \subset A_2 \subset A_0$), because PVR and PTR were needed in studies on the identification of unusual data [V, VI]. CYM was known to be an important parameter for the identification of the small motor urge incontinence class on the basis of medical knowledge of the expert physicians [I]. Paper (VI) showed that CYM was indeed a useful attribute: CYM was often included into the automatically generated diagnostic rules.

The attributes seemed to be insufficient for successful unsupervised classification. Hierarchical cluster analysis [Eve93] detected clusters corresponding only to the small normal class and was unable to clearly separate the larger incontinence classes [I]. The *k*-means algorithm [Eve93] gave better results [LJP97+, III], but supervised classifiers outperformed this method [III]. Since unusual data often makes unsupervised classification difficult, we studied in [LJ01] clustering of the data set D_0' which was cleaned by removing noise and outliers. Some hierarchical methods managed to produce clusters that corresponded to the diagnostic classes, but the overall result was that hierarchical clustering could not separate classes even in the cleaned data. However, *k*-means algorithm performed quite well with the cleaned data.

5.2.3. Identifying the appropriate proximity function

When we started the research of data pre-processing with instance-based methods, the first consideration was to find a proximity measure suitable for mixed data with missing values. The problem was that most of the proximity functions assume the same level of measurement for all attributes. Many real world data sets, however, have attributes of mixed type. A proximity function designed for one type of data may not be the best choice, especially for mixed data with nominal attributes. For example, Euclidean distance treats nominal attributes, whose values do not have meaningful order, as if they were quantitative.

In paper (IV) we compared Manhattan and Euclidean distances - the two variants of Minkowskian distance function, with three heterogeneous proximity functions, which treat the attributes differently according to their scales, to find out if there were truly differences between the functions. The heterogeneous proximity functions were Gower's similarity function [Eve93], Aha's heterogeneous Euclidean-overlap metric (HEOM) [AKA91, WM97], and heterogeneous value difference metric (HVDM)

[WM97]. We defined the scales in this comparative study as in [WM97], where ordinal attributes were treated as quantitative attributes. Missing values caused the distance between the values to be maximal, i.e. 1, for all types of attributes in the Minkowskian distance functions, HEOM, and HVDM [WM97]. Instead of this 'pessimistic' approach, Gower's similarity function used 'ignore' strategy, where attributes with missing values were not considered in the proximity evaluation [Eve93].

Our experiments showed that a heterogeneous proximity function is not necessarily better than a proximity function assuming the same scale for the data. The differences in accuracies were mainly insignificant and the Minkowskian functions outperformed Gower's similarity function and HEOM. However, significant differences ($p < 0.05$) were in favour of HVDM, which treats the nominal attributes more appropriately than the other functions studied. The significant differences in TPRs favoured HVDM, and TPRs of the other functions behaved as the accuracies: The Minkowskian distance functions outperformed Gower's similarity function and HEOM. Our results concerning the relative performance of the Euclidean, HEOM, and HVDM were in accord with those of Wilson *et al.* [WM97].

There were five reasons for performing the study in paper (IV) although Wilson *et al.* [WM97] showed with an extensive comparison that HVDM is a better choice for mixed data than the Euclidean distance or HEOM. First, we wanted to test how TPR would behave, because only accuracies were studied in [WM97]. Since accuracy has its shortcomings (see Section 4.6), TPR should also be examined. Second, we felt that the possibility of identifying significant differences by chance in multiple comparisons should be taken into account [Pet97, Sal99]. Bonferroni correction was made to accommodate the potential for increased Type I error [Pet97, Sal99]. Third, due to small sample sizes a nonparametric paired test was in our opinion more appropriate that the paired t test used in [WM97]. Fourth, 90% confidence level ($p < 0.1$) applied in [WM97] is rather low, and, therefore, we used the higher 95% confidence level ($p < 0.05$). Last, we did not want to rely only on the data retrieved from the UCI machine learning repository [BM98], because we are specifically concerned with medical data, and, moreover, it is sometimes risky to assume that the UCI data represent the actual real world situation [SN98]. For these reasons, seven of our real world biomedical data sets were included in the study.

Although experimental work in [WM97, IV] showed that HVDM is a good choice for the mixed data, the metricity of this distance function is an important issue. Since Wilson *et al.* [WM97] did not study whether HVDM or other of their new distance functions are metric, we have proven the metricity of HVDM in [JL01].

5.2.4. Balancing the imbalanced class distribution

In paper (V), we presented a new instance-based method for balancing imbalanced class distribution before the data analysis. A more detailed description of the method is found in [Lau01]. The neighbourhood cleaning rule (NCL) is based on the idea of the one-sided selection (OSS) method by Kubat *et al.* [KM97] which is an instance-based data reduction method for reducing the larger class when the class distribution of a two-class problem is imbalanced. NCL utilises the OSS principle, but considers more carefully the quality of the data to be removed. The major drawback of OSS is that the data reduction process is quite sensitive to noise. Although noise is removed after data reduction, the result is not the best possible because of a large amount of noise in the remaining data. Moreover, noise is usually removed before statistical analyses and data mining.

The basic idea of our method is the same as in OSS: All examples in the class of interest C (positive class) are saved, while the rest O of the original data T (negative classes) is reduced. NCL can be applied to several classes of interest, but for the sake of clarity, we discuss here only the setup of one class against the other classes. In contrast to OSS, NCL emphasizes more data cleaning than data reduction. Our justification for this approach is two-fold. Firstly, the quality of classification results does not necessarily depend on the size of the class. There are small classes that can be identified easily and large classes that are difficult to classify. Therefore, besides the class distribution, we should consider other characteristics of data, such as noise, that may hamper classification. Secondly, studies of data reduction with instance-based techniques [WM00] have shown that it is difficult to maintain the original classification accuracy while the data are being reduced. This aspect is important, since while improving the identification of small classes, the method should be able to classify other classes with acceptable accuracy.

Consequently, we chose Wilson's edited nearest neighbour rule (ENN) [WM00] to identify noisy data A_1 in O. ENN removes examples whose class differs from the majority class of the three nearest neighbours in O. ENN retains most of the data,

while maintaining a good classification accuracy [WM00]. In addition, we cleaned neighbourhoods of examples belonging to C: The three nearest neighbours in T that misclassify examples of C and are members of O are inserted into the set A_2. To avoid excessive reduction of small classes, only examples from classes larger than or equal to $0.5 \cdot | C |$ are considered while forming A_2. Lastly, the union of sets A_1 and A_2 is removed from T to produce the reduced data set S. To make NCL to suit better for solving real world problems than OSS, we utilised HVDM and designed NCL with multi-class problems in mind. Our method was named the neighbourhood cleaning rule, because it considers data cleaning in neighbourhoods from two viewpoints. Negative classes are cleaned by using neighbourhoods for noise removal, whereas the neighbourhoods that misclassify examples of the class of interest are removed. Algorithm 1 shows the functioning of the NCL method.

Algorithm 1: Neighbourhood cleaning rule.

Input: Original data T, Index i of the class of interest.
Output: Reduced data S.

1. Split T into the class of interest C_i and the rest of data O with classes C_j ($i \neq j$).
2. Identify noisy data A_1 in O with the edited nearest neighbour rule.
3. For each case $x \in C_i$
   if (x is misclassified by its 3-nearest neighbours Y in T)
      for each $y \in Y$
         if ($y \in C_j$) and ($| C_j | \geq 0.5 \cdot | C_i |$) then $A_2 = \{ y \} \cup A_2$
4. $S = T - (A_1 \cup A_2)$

NCL outperformed simple random sampling within classes and the OSS method in the experiments with ten complete data sets that had a small and difficult class. All reduction methods clearly improved identification of these classes (20-30%), as measured with the mean TPR of the three-nearest neighbour method and C4.5 decision tree generator, but differences between the methods were insignificant ($p < 0.05$). However, the significant differences in accuracies, TPRs and TNRs obtained from the reduced data were in favour of NCL. The results suggest that NCL is a useful

method for improving modelling of difficult small classes, as well as for building classifiers that identify these classes from real world data which often have an imbalanced class distribution.

5.3. Galactica - a genetics-based machine learning system

The strengths of genetic algorithms and research indicating that machine learning systems based on genetic algorithms can solve medical problems at least as well as the conventional systems [BP91, Jan93, Con95] motivated us to develop a general purpose learning system, named Galactica [II], that utilises genetic algorithms. Another motivation was the success of ANNs, which are also inspired by nature, in solving medical problems [FD95, PEJ96, Cha98].

Galactica is based on the ideas presented in GABIL [DJSG93] and GIL [Jan93] systems. Our method resembles more the straightforward GABIL than GIL. Both GABIL and Galactica utilise the standard 'off the shelf' genetic algorithm with minor modifications, while GIL is a complex system that makes use of many task specific learning operators. The major difference in the design of the GABIL and Galactica systems is the presentation of the examples to the learning system. GABIL learns in a batch-incremental manner: One or few examples are presented to the system at a time, and if the current rules misclassify the new data, the rules are updated using all the training data. On the other hand, Galactica is a batch learner that uses all the training data on the each iteration of the genetic algorithm. In addition, completeness, consistency, and complexity of the rule sets are presented as separate terms in the fitness function of Galactica.

Galactica uses a simple genetic algorithm to learn rules for two-class problems inductively and in a supervised manner from examples which are characterised by categorical attributes. The genetic algorithm maintains a population of variable-length chromosomes from which the fittest individuals are selected for reproduction. The parents are modified with the crossover and mutation operators to generate a new generation. The populations are non-overlapping populations which are selected by the roulette wheel method [Gol89]. Elitism is used, i.e., the best member of each generation is moved intact into the new generation. Two of Michalski's concept learning operators, adding condition and dropping conditions [Mic83, Jan93], are used to improve the learning process. After the genetic algorithm has terminated, chromosomes can be decoded as symbolic rules for examination.

Knowledge is represented in sub-symbolic chromosomes for the genetic algorithm and in symbolic IF-THEN type decision rules for humans. For example, a rule for the stress incontinence diagnosis might be: *IF [Urine in vagina = No] & [Stress symptom = Yes] & [Difficulties with voiding = No] & [Urge symptom = No] THEN Stress incontinence*. The chromosomes are coded as binary strings so that for each condition as many bits are reserved as the corresponding attribute has values, and the condition bit is set at one if the attribute has a corresponding value, otherwise the bit is set at zero [DJSG93, Jan93, YZ96]. If an attribute does not exist in a rule, all of its value bits are turned to one, i.e. the attribute can have any value and is therefore meaningless.

The fitness of a chromosome is mostly based on the number of positive and negative examples covered in the training set. The fitness increases as the positive cover grows and the negative cover diminishes. In addition, the complexity of the chromosome slightly affects its fitness; the simpler chromosomes are considered better than the more complex ones. A binary coded example is covered by the chromosome when the logical AND operation between the example and the chromosome returns the example unchanged. In other words, the example is covered if the attribute values both in the example and in the chromosome match. Missing values do not prevent classification, because match is considered to occur when an attribute is absent from the example or chromosome, or when the attribute is absent both from the example and chromosome.

5.4. Comparison of Galactica and other classification methods

Galactica was compared in paper (III) to discriminant analysis [Sha96], logistic regression [Agr96], *k*-means cluster analysis, C4.5 decision tree generator, and a random bit climber (RBC) [Dav91b]. The methods were evaluated in the diagnosis of female urinary incontinence in terms of prediction accuracy of classifiers produced from the patient data. The missing data were imputed with rounded means to allow application of the statistical methods. The classification ability of machine learning methods was also compared in the original incomplete data. The task was to classify stress, mixed, and sensory urge cases by using the principle of one class against the other classes.

Discriminant analysis, logistic regression, C4.5, and Galactica produced the most accurate classifiers from the imputed data D_0' (mean prediction accuracy 89-91%).

The statistical differences in the prediction accuracy of these classification methods were insignificant. RBC performed slightly worse than these methods and *k*-means cluster analysis was clearly the weakest method. The poor performance of the *k*-means algorithm was expected, because the hierarchical clustering methods could not recover diagnostic classes from the data in [I]. C4.5 was the best, Galactica the next best, and RBC was the worst method with the original incomplete data. The mean prediction accuracies of these methods ranged from 84% to 92%. The unexpectedly good performance of RBC suggests that the classification task was such that the more advanced methods were not able to express their full power. However, the results were in agreement with the results of earlier research indicating that genetic algorithms are a competitive method for constructing classifiers from medical data.

Our system was applied in [KLP99+] to discover rules from the data on otological diseases involving vertigo [Ken96]. The vertigo data had 564 examples that were described with the subset of 38 mixed attributes of all the 170 available attributes. Galactica learned diagnostic rules, with the principle of one class against the other classes, for vestibular schwannoma (N = 128), benign paroxysmal positional vertigo (N = 59), Menière's disease (N = 243), sudden deafness (N = 21), traumatic vertigo (N = 53), and vestibular neuritis (N = 60) diagnoses. The prediction accuracies (and true positive rates) of rules for these diagnoses were 91% (62%), 96% (74%), 81% (76%), 95% (11%), 92% (28%), and 98% (90%) respectively. Besides being accurate, the rules contained the five most important diagnostic questions identified in the earlier research [Ken96]. True positive rates indicated that our method had difficulties in identifying the small sudden deafness and traumatic vertigo classes. In a comparison of the accuracies obtained with discriminant analysis and genetic algorithms [KLJ00+], discriminant analysis outperformed Galactica slightly. These results showed that our method could also solve a larger classification problem accurately.

Galactica, C4.5, and the three-nearest neighbour method (3-NN) were compared indirectly in [VI], where different methods were used to construct an ensemble classifier from the EM imputed data D_0'. The mean prediction accuracies of 11 classifiers built with these methods were 83%, 85%, and 85% respectively. The accuracies were lower than in paper (III), because the multiclass situation is more difficult for learning algorithms than the setup of one class against the others. Similar to RBC, the simple 3-NN was able to produce results comparable with the results of

the more complex methods. On the other hand, 3-NN with HVDM seems to be as good as C4.5 rules produced with default settings [V].

We considered machine learning systems C4.5 and Galactica preferable for the automatic construction of medical decision aids, because they can cope with missing data values directly and can present a classifier in a comprehensible form. For these reasons they were used in the expert system [VI]. The presentation of knowledge in a comprehensible manner is crucial in areas such as decision making and medicine, where humans must fully understand the classifiers [HW90, MK90, FD95]. Logistic regression and discriminant analysis create a mathematical model which gives valuable information, for example, about the dependencies between diagnostic parameters and diagnosis. Understanding and thus correct interpretation of these non-symbolic models is usually quite difficult for individuals who do not have considerable statistical knowledge. Consequently, medical decision support systems constructed from the results of statistical methods are black boxes having limited capabilities to explain their decisions. Moreover, rigorous evaluation and testing may be far more difficult than with transparent systems [HW90]. Conversely, decision trees and rules are understandable without extensive expertise. An expert in the problem area, for example a physician, can directly evaluate and verify this type of classifier provided that it is not too complex.

ANNs have been applied successfully in decision support in medicine [FD95, Cha98]. Artificial neural networks were not included in the comparison of paper (III), because we suspected that there were not enough data for the network [III]. In the diagnosis of acute abdominal pain, for example, Pesonen *et al.* [PEJ96] used a neural network with 14 input nodes, two output nodes, and 1333 examples, while in [III] we used 13 attributes, two classes, and had only 529 examples available. The lack of data is a common problem in medical applications of ANNs, because an abundance of examples is needed to train and validate a neural network, but it is often impossible to collect large medical data sets [FD95]. Another reason why we did not strive for ANNs was their sub-symbolic knowledge representation, which makes ANNs essentially black boxes. Interpreting of the knowledge captured in the weights of the neural network is even more difficult than understanding the decision models built with statistical methods. Moreover, ANNs can process only complete data, which is quite rare in the area of medicine.

We recently tested a feedforward multilayer perceptron using the attributes in the set A_1, excluding age and including CYM, as the input nodes. The network had four hidden nodes and three output nodes that corresponded to the stress, mixed, and the other female urinary incontinence diagnoses. The network was trained with the EM imputed data set D_0', backward propagation algorithm, and 10-fold cross-validation. The average accuracy and TPRs for the three classes were 78%, 88%, 65%, and 51% respectively. As we anticipated in (III), ANNs were not suitable for this data set: the classification ability was moderate, the diagnostic classes had to be combined, and not all the available attributes could be used. Better results might be obtained with a larger data set.

5.5. Female urinary incontinence expert system

Paper (VI) describes IES1 which is an Internet-based multi-expert system. Multi-expert refers to the ensemble classifier that was constructed from 3-NN classifier, and from decision rules produced with Galactica and C4.5. A total of 11 classifiers was built from the EM imputed D_0' by using a 70:30% ratio split into the training and testing sets and attribute set A_2. Diagnostic rules were crafted manually for the rare overflow urinary incontinence and fistula diagnoses. Training data for 3-NN was reduced with NCL to improve identification of the small and difficult sensory urge class.

The original aim was to use an ensemble classifier EN0 that would give a single diagnosis based on the majority voting scheme. It was assumed that the different learning biases and different knowledge presentations would result in classifiers that make errors in different but complementary manner. Unfortunately, EN0 classifier performed approximately as well as the individual classifiers. However, the ensemble classifier EN1, which gives from one to three diagnoses, was clearly better than the individual classifiers and EN0. The good performance of EN1 does not stem simply from its multi-diagnosis feature. On average, 20% of the EN1 classifications were multiple diagnoses, and of these on average 6% included three different diagnoses. EN1 improved EN0 classifications especially because it could correctly identify more mixed and sensory urge cases than EN0.

$EN1_{med}$, an EN1 classifier with median accuracy of 11 EN1 classifiers, was used to implement the inference module of IES1. The rationale for building a number of classifiers and then using a mediocre classifier instead of the best one was the need for

objectivity. It is known that one can usually generate a good model or classifier for a data set by merely building enough of them [HMS01] and, therefore, a data miner should try to avoid this pitfall, for example, by using a classifier that performs moderately. The first diagnosis of $EN1_{med}$ was correct in 86% of the test cases in 'laboratory conditions'. When the second and the third diagnoses were considered, the classifier correctly identified 94% of the test data of D_0', the medians and the ranges of TPRs and TNRs being 97%, 82-100% and 94%, 87-96% respectively. The accuracy of $EN1_{med}$ was 91% in the data set D_1. If the misclassification of the only normal case is ignored, the medians and the ranges of TPRs and TNRs in D_1 were 96%, 70-100% and 90%, 75-95% respectively.

We generated 'exception' rules from the noisy cases identified with NCL and outlier thresholds [LJK00] for individual attributes. The system comments on request whether a case which a physician has diagnosed as stress or mixed incontinence is unusual. A case may have unusual value combinations or a quantitative attribute, for example MUC, may have unexpectedly low or high values. We expect that this function draws the physician's attention to the possible errors in individual values and to possibly misclassified or borderline cases, but the final usefulness and benefit of this feature remains to be seen. Our approach is opposite to the work in [FNI91], where decision trees were built from the reliable cases to identify unusual thyroid patients. We also considered this approach, but since we did not have reliability information available, 'exception' rules were generated.

The expert system was implemented as a Java applet that a physician can access via Internet with a Java enabled World Wide Web (WWW) browser (see Figure 4). The client-server architecture that IES1 utilises has some advantages over the stand-alone expert systems. Firstly, the user does not have to worry about the software installation or maintenance, which is sometimes problematic in the case of the traditional stand-alone applications. All that is needed is a relatively new Java-enabled WWW browser, such as Netscape Navigator or Microsoft Internet Explorer. Secondly, Java applets have continuous running state, which makes the system like a normal application for a user, except that it must be launched from a browser. The expert system can interact with the user and locally perform tasks, such as error checking and user-interface manipulation, which are appropriate there. Lastly, Java offers a platform independent way to use our system.

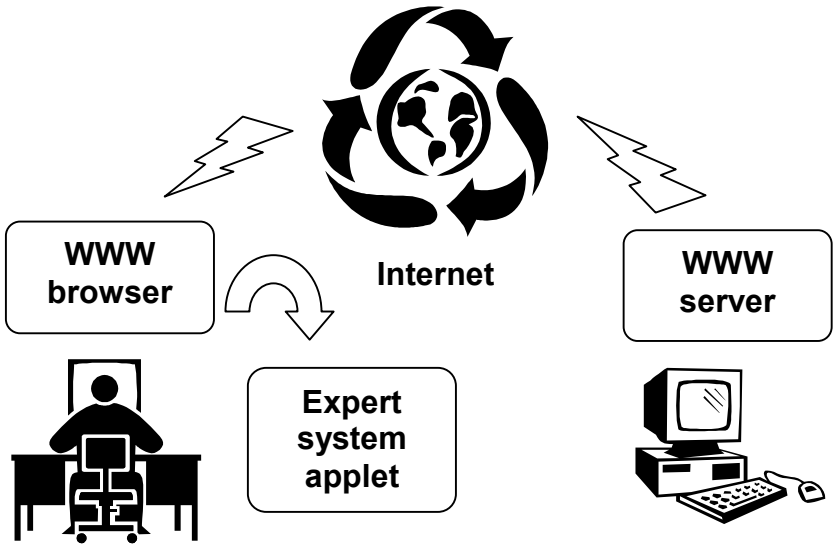

Figure 4. The architecture of IES1. The expert system is implemented as a Java applet that a physician may load into a WWW browser over Internet.

6. Discussion and conclusions

The aims of this work were both practical and scientific. The practical aim was to develop an expert system as a decision support tool for the differential diagnosis of female urinary incontinence. The main scientific aim was to study whether data mining techniques could be used as an alternative to the conventional knowledge acquisition process, where a knowledge engineer extracts knowledge from an expert via lengthy interviews. Pre-processing of the data and discovery of the diagnostic knowledge with machine learning methods were paid special attention. This work produced a new machine learning program, Galactica, which is based on genetic algorithms, and the neighbourhood cleaning rule (NCL) that balances the imbalanced class distribution by using an instance-based approach. Galactica, NCL, and other data mining techniques were applied to overcome difficulties of real world data, and to develop ability for classification and critique for the Internet-based incontinence multi-expert system IES1.

Galactica proved to be a competitive approach for building classifiers in comparison with different machine learning and statistical methods in two real world medical data sets. In female urinary incontinence data the system was among the best, and in data of otological diseases involving vertigo the system performed approximately as well as the discriminant analysis. Genetic algorithms are able to roam efficiently large search spaces without being trapped in a local minimum or maximum [Gol86]. This kind of search would be extremely useful in data mining where search spaces are huge. Unfortunately, genetic algorithms are computationally complex and, consequently, require a lot of processing time. However, we believe that the genetic algorithm paradigm in machine learning is worth pursuing, but our method needs further research to be applicable to the mining of large databases. The system should also be developed to solve multi-class problems with mixed attributes. In addition, a carefully planned comparison with other methods by using a number of the UCI data sets as well as our real world data sets is needed to ascertain Galactica's performance in various different tasks.

Our contribution to the pre-processing of real world data was a new instance-based method NCL that improves identification of difficult small classes by balancing the imbalanced class distribution with data reduction. NCL outperformed simple random

sampling within classes and the OSS method in the experiments with 10 data sets. Improvement in TPRs of the class of interest was 20-30%, but the differences between the methods were insignificant. However, NCL was more successful than other methods in maintaining the original classification ability of the other classes. Our results suggest that NCL is a useful method for improving descriptive analysis of difficult small classes. NCL may also be used for building classifiers that better identify difficult small classes from real world data, which frequently have an imbalanced class distribution. The drawback of the method is its complexity, because the building of the proximity matrix takes $O(N^2)$ time, where N is the number of examples.

Our studies showed that it is possible to develop an expert system by making use of knowledge discovered with machine learning. In addition, data pre-processing results allowed us to implement functionality that draws users' attention to possible errors or borderline cases. The expert system not only aids physicians in the diagnosis of the incontinent women, but it also acts as a critic for the physicians' decisions [GW98]. The diagnostic ability of the system was quite good. The first diagnosis was correct in 86% of the test cases in 'laboratory conditions'. When the second and the third diagnoses were considered, IES1 correctly identified 94% of the test data of D_0', while 91% of the data set D_1 were correctly classified. The Internet-based client-server architecture allows access and distribution of the expert system through Java-enabled WWW browsers. Physicians are currently evaluating the usefulness of IES1 in diagnostic work. Although the classification ability of the system is good, only prolonged field-use can qualify an application developed with machine learning as a successful real world system [SN98]. Even though the symptoms of stress are sometimes so unmistakable that surgery may be carried out without urodynamic measurements, previous research suggests that diagnoses based solely on the symptoms may not be accurate enough [JNO94]. The real world diagnostic work-up will show whether IES1 is able to reduce the number of expensive urodynamic measurements.

Patient data collected in a local hospital may not be a representative sample of the population and, moreover, different clinical procedures in different organisations may further bias the data. Therefore, future research should explore methodologies to make IES1 easily applicable in different hospitals. One approach could be to discover diagnostic knowledge from a data set created by combining locally collected data sets,

but this approach would require carefully planned and extensive co-operation between various organisations. Locally built classifiers might be a better approach to adapt the classifiers to the differences in the procedures and patients of the different organisations. Furthermore, ontologies might be of great help in the development of generally applicable and easy to use DSSs for the differential diagnosis of the female urinary incontinence. Another area for future research is the planning of the therapy for the incontinent women. The methodologies used to develop diagnostic capability for IES1 may be used to build classifiers for therapy as well.

One can justifiably ask whether IES1 is really an expert system. It is difficult to answer this question, because the terms DSS, KBS, and expert system are nowadays used quite freely. For example, the expert system by Güvenir *et al.* [GE00] for the diagnosis of erythemato-squamous diseases is similar to IES1, but uses only sub-symbolic knowledge representation. Likewise, Chae [Cha98] makes no distinction between the traditional rule-based systems and sub-symbolic ANNs in his review of medical expert systems. IES1 is, as a computer program designed to aid physicians, definitely a DSS. Clearly, domain knowledge of IES1 is captured in symbolic rules produced by Galactica and C4.5, and in a function produced by 3-NN that maps classes to examples. IES1 does not have a traditional inference engine that controls the reasoning, because its rules are simple classification rules, but the rule matching and voting features of the system may be seen as the general problem solving knowledge. In addition, the diagnostic knowledge is mostly represented in symbolic rules which explain the reasoning behind the automatic diagnosis to a user. For these reasons, IES1 is in our opinion a KBS. The ability to perform as well as a human expert in a specific task makes a KBS an expert system (see Section 2.2). Unfortunately, we have neither found any comparison of the diagnostic accuracy of a physician and a computer program in diagnostic work-up of female urinary incontinence, nor any comprehensive study of the accuracy of physicians' diagnoses in this area of medicine. However, we argue that IES1 probably meets this requirement, because its classification ability was over 90% and TPRs were 70-100%. This is quite a good result for a human or a computer when the only data available are examples.

The greatest drawback of using EIE learning is that the discovered knowledge is likely to be shallow. For example, the rules of IES1 are straightforward decision rules whose THEN part simply gives the predicted class (see Section 5.3). These rules do

not allow inference chains that are traditionally applied to mimic expert's reasoning. However, it should be noted that the terms deep and shallow knowledge are extremes, and in practice the knowledge captured in expert systems is somewhere between the extremes. There probably exist many manually crafted expert systems whose knowledge is partly shallow due to the difficult knowledge acquisition process. Likewise, knowledge produced with the EIE learners is not totally shallow; it may capture some aspects of deep knowledge. Expert system development with machine learning methods also has problems caused by real world data, such as those we encountered during building IES1. However, we believe that these difficulties are smaller than difficulties related to the manual knowledge acquisition. Maintaining of the knowledge and the different approaches to solve problems in different organisations are problematic for the both development approaches.

The usefulness of automatic knowledge discovery with machine learning depends largely on the aims of a data miner and the problem characteristics. If the aim is to analyse and model the data, automatically acquired knowledge is usually of great value to the domain experts. The situation is more complex when the knowledge is further utilised in a DSS aimed for the classification of data. The developers must decide whether the knowledge needs additional refinement and whether the explanations based on the knowledge are sufficient for a user. Symbolic EIE learners may be considered as preliminary tools that are used to extract raw knowledge for further use in expert system shells, where knowledge is rearranged and cleaned to produce the final expert system [Tur93 Chapter 13.12]. A more straightforward approach is to use knowledge 'as it is' only with slight modifications needed to utilise it. IES1, the expert system by Güvenir *et al.* [GE00], and many ANNs based expert systems are examples of this approach. For example, in IES1 a class hierarchy, with the base class 'Classifier', was designed, and individual classifiers were implemented as subclasses whose methods were overridden. If the knowledge is symbolic, expert systems built with machine learning methods may provide explanations for decisions, although explanations are more limited than in the traditional expert systems. If explanations are not needed, sub-symbolic EIE learners, such as ANNs, are applicable. The use of machine learning may be the best or the only alternative for knowledge acquisition if the problem is such that experts are not available, the expert is unable to verbalise his or her knowledge, or the problem is so complex that it is almost impossible for a human to understand. In these cases machine learning is of

great help for humans who require decision support and possibly need to understand the problem.

If partial or missing explanations are acceptable, expert systems based on knowledge obtained with the EIE learners are a viable alternative to manually built expert systems. The machine learning approach is inexpensive and fast in comparison with manual expert system development providing that the data are available or that data collection is not overly slow or difficult. Waterman [Wat86] estimates that a moderately difficult expert system project with a 2-4 person team requires 1-6 person years, while a very difficult task with 4-6 people takes 10-30 person years. In comparison, C4.5 often discovers knowledge in few seconds or minutes. Although the machine learning approach is clearly more automatic than manual and semi-automatic knowledge acquisition, in practice it is not a fully automatic acquisition method [Tur93 Chapter 13.5]. In fact, quite a lot of interaction is sometimes needed between a data miner, who replaces a knowledge engineer, an expert, and a user who is often an expert. It is especially important to understand a user's problem and that the user participates actively in the application process [SN98]. The interaction between different parties takes time, but since the knowledge discovery is automated, the total development time, including the time needed to put knowledge into operation, will be only a fraction of the time required with the conventional approach [GSB93+, WWZ99].

According to Hill *et al.* [Tur93 p. 537], the ideal knowledge acquisition system should have the following characteristics: (1) direct interaction with expert without knowledge engineer, (2) applicability to unlimited problem domains, (3) tutorial capabilities, (4) detection of inconsistencies and gaps in knowledge, (5) ability to use many sources of knowledge, (6) easy user interface, and (7) ability to inference with different expert system tools. We doubt that any system will ever fully meet all these requirements, and agree with Turban [Tur93 Chapter 13.12] in that machine learning is still limited in its capabilities in the area of knowledge acquisition. However, the most important requirements (1) and (2) are met. Machine learning can bypass lengthy interviews and is applicable in various problem areas. The ensemble learning approach applied in IES1 addresses characteristic (5). Although automatic knowledge acquisition is still in progress, several authors agree with us in that the knowledge discovery with machine learning is one of future areas in the field of the expert systems [H-RJ94, DeH98, Dur98, Alt99]. There is also evidence to back this claim.

Expert systems described in [GSB93+, BD-M95, Tsu98, WWZ99, CFC01+] have proven to be as good as human experts or manually crafted expert systems. Another promising trend is the use of WWW in connection with expert systems [Dur98]. We also addressed this new area in IES1.

Future research might focus on the development of an expert system building tool that utilises multiple EIE learners and other data mining methods to automatically extract knowledge from data. This type of system would generate automatically or interactively classifiers, ability for critique, and explanations for decisions for an expert. In addition, the system might automatically code this functionality into the rules of an expert system shell or the system might produce a Java applet or other Internet-based application that could be used both locally and via Internet. In addition, the available ontologies might be used automatically in the expert system development. This approach might enable an expert to build expert systems directly without the need to consult neither knowledge engineer nor a data miner.

To summarise, we constructed an Internet-based multi-expert system for the differential diagnosis of female urinary incontinence by using data mining techniques. Inspired by the practical problem, we developed methods for machine learning and data pre-processing. The genetics-based machine learning system Galactica showed to be a competitive method for classifying medical data. The neighbourhood cleaning rule outperformed the other two methods for balancing the imbalanced class distribution. Although further work is needed, automatic knowledge discovery seems to be a good alternative for the manual knowledge acquisition in the expert system development.

References

[Agr96] Agresti A: *An Introduction to Categorical Data Analysis*, Wiley, New York, 1996.

[AKA91] Aha DW, Kibler D, Albert MK: Instance-based learning algorithms, *Machine Learning* 6 (1991) 37-66.

[Alt99] Altman RB: AI in medicine: The spectrum of challenges from managed care to molecular medicine, *AI Magazine* 20:3 (1999) 67-77.

[AN95] Aamodt A, Nygård M: Different roles and mutual dependencies of data, information, and knowledge - An AI perspective on their integration, *Data & Knowledge Engineering* 16 (1995) 191-222.

[BC83] Beckman RJ, Cook RD: Outliers, *Technometrics* 25 (1983) 119-149.

[BD-M95] Ben-David A, Mandel J: Classification accuracy: Machine learning vs. explicit knowledge acquisition, *Machine Learning* 18 (1995) 109-114.

[BFM97] Bäck T, Fogel DB, Michalewicz Z: *Handbook of Evolutionary Computation*, Institute of Physics Publishing and Oxford University Press, Bristol, 1997.

[BL87] Barnett V, Lewis T: *Outliers in Statistical Data*, Wiley, Norwich, 2nd edn., 1987.

[BL97] Blum AL, Langley P: Selection of relevant features and examples in machine learning, *Artificial Intelligence* 97 (1997) 245-217.

[BM98] Blake CL, Merz CJ: *UCI Repository of machine learning databases*, Irvine, University of California, Department of Information and Computer Science, 1998. [http://www.ics.uci.edu/~mlearn/MLRepository.html]

[BP91] Bonelli P, Parodi A: An efficient classifier system and its experimental comparison with two representative learning methods on three medical domains, In: Belew RK, Booker LB (eds.): *Proceedings of the Fourth International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, 1991, pp. 288-295.

[CFC01+] Chapman WW, Fizman M, Chapman BE, Haug PJ: A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia, *Journal of Biomedical Informatics* 34 (2001) 4-14.

[Cha98] Chae YM: Expert systems in medicine, In: Liebowitz J (ed.): *The Handbook of Applied Expert Systems*, CRC Press, Boca Raton, 1998, pp. 32-1 - 32-20.

[Con95] Congdon CB: *A Comparison of Genetic Algorithms and Other Machine Learning Systems on a Complex Classification Task From Common Disease Research*, Ph.D. Thesis, Department of Computer Science and Engineering, University of Michigan, 1995.

[Dav91a] Davis L: *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York, 1991.

[Dav91b] Davis L: Bit-climbing, representational bias, and test suite design. In: Belew RK, Booker LB (eds.): *Proceedings of the Fourth International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, 1991, pp. 18-23.

[DeH98] De Hoog R: Methodologies for building knowledge based systems: Achievements and prospects, In: Liebowitz J (eds.): *The Handbook of Applied Expert Systems*, CRC Press, Boca Raton, 1998, pp. 1-1 - 1-14.

[DJSG93] De Jong KA, Spears WA, Gordon DF: Using genetic algorithms for concept learning, *Machine Learning* 13:2-3 (1993) 161-188.

[DL97] Dash M, Liu H: Feature selection for classification, *Intelligent Data Analysis* 1 (1997) 131-156.

[DLS72+] De Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC: Computer-aided diagnosis of acute abdominal pain, *British Medical Journal* 2 (1972) 9-13.

[Dur96] Durkin J: Expert systems: A view of the field, *IEEE Expert* 11:2 (1996) 56-63.

[Dur98] Durkin J: Expert system development tools, In: Liebowitz J (ed.): *The Handbook of Applied Expert Systems*, CRC Press, Boca Raton, 1998, pp. 4-1 - 4-26.

[Eve93] Everitt BS: *Cluster Analysis*, Arnold, London, 3rd edn.,1993.

[FD95] Forsström JJ, Dalton KJ: Artificial neural networks for decision support in clinical medicine, *Annals of Medicine* 27 (1995) 509-517.

[FNI91] Forsström JJ, Nuutila P, Irjala K: Using the ID3 algorithm to find discrepant diagnoses from laboratory databases of thyroid patients, *Medical Decision Making* 11 (1991) 171-175.

[FP-SS96] Fayyad U, Piatetsky-Shapiro G, Smyth P: From data mining to knowledge discovery in databases, *AI Magazine* 17:3 (1996) 37-54.

[GE00] Güvenir HA, Emeksiz N: An expert system for the differential diagnosis of erythemato-squamous diseases, *Expert Systems with Applications* 18 (2000) 43-49.

[GMP96+] Glymour C, Madigan D, Pregibon D, Smyth P: Statistical inference and data mining, *Communications of the ACM* 39:11 (1996) 35-41.

[Gol89] Goldberg DE: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, 1989.

[Gol94] Goldberg DE: Genetic and evolutionary algorithms come of age, *Communications of the ACM* 37:3 (1994) 113-119.

[Gor95] Gorman R: Expert system for management of urinary incontinence in women, In: *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, AMIA Inc., 1995, pp. 527-531.

[GSB93+] Giordana A, Saitta L, Bergadano F, Brancadori F, De Marchi D: ENIGMA: A system that learns diagnostic knowledge, *IEEE Transactions on Knowledge and Data Engineering* 5:1 (1993) 15-28.

[GW98] Gertner AS, Webber BL: TraumaTIQ: Online decision support for trauma management, *IEEE Intelligent Systems* 13:1 (1998) 32-39.

[HHH98+] Hunt DL, Haynes RB, Hanna SE, Smith K: Effects of computer-based clinical decision support systems on physician performance and patient outcomes: A systematic review, *Journal of the American Medical Association* 280:15 (1998) 1339-1346.

[HK01] Han J, Kamber M: *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2001.

[HMS01] Hand D, Mannila H, Smyth P: *Principles of Data Mining*, MIT Press, 2001.

[H-RJ94] Hayes-Roth F, Jacobstein N: The state of knowledge-based systems, *Communications of the ACM* 37:3 (1994) 27-39.

[HSB94+] Hernández C, Sancho JJ, Belmonte MA, Sierra C, Sanz F: Validation of the medical expert system RENOIR, *Computers and Biomedical Research* 27 (1994) 456-471.

[Hu90] Hu TW: Impact of urinary incontinence on health-care costs, *Journal of American Geriatrics Society* 38 (1990) 292-295.

[HW90] Hart A, Wyatt J: Evaluating black-boxes as medical decision aids: Issues arising from a study of neural networks, *Medical Informatics* 15 (1990) 229-236.

[Jan93] Janikow CZ: A knowledge-intensive genetic algorithm for supervised learning, *Machine Learning* 13:2-3 (1993) 189-228.

[JD88] Jain AK, Dubes RC: *Algorithms for Clustering Data*, Prentice-Hall, New Jersey, 1988.

[JL01] Juhola M, Laurikkala J: On metricity of heterogeneous Euclidean-overlap metric and heterogeneous value difference metric with missing values, *Pattern Recognition*. (manuscript)

[JNO94] Jensen JK, Nielsen FR, Ostergard DR: The role of patient history in the diagnosis of urinary incontinence, *Obstetrics and Gynecology* 83:5 (1994) 904-910.

[KAJ98+] Kentala E, Auramo Y, Juhola M, Pyykkö I: Comparison between diagnoses of human experts and a neurotologic expert system, *Annals of Otology, Rhinology & Laryngology* 107:2 (1998) 135-140.

[Ken96] Kentala E: Characteristics of six otologic diseases involving vertigo, *The American Journal of Otology* 17 (1996) 883-892.

[KHM98] Kubat M, Holte RC, Matwin S: Machine learning for the detection of oil spills in satellite radar images, *Machine Learning* 30 (1998) 195-215.

[KLJ00+] Kentala E, Laurikkala J, Juhola M, Pyykkö I: Comparison of diagnostic accuracy between genetic algorithm and discriminant analysis in neurotologic expert system, In: Claussen CF, Haid CT, Hofferberth B (eds.): *Equilibrium Research, Clinical Equilibriometry and Modern Treatment*, Elsevier, 1999, pp. 669-673.

[KLP99+] Kentala E, Laurikkala J, Pyykkö I, Juhola M: Discovering diagnostic rules from a neurotologic database with genetic algorithms, *Annals of Otology, Rhinology & Laryngology* 108:10 (1999) 948-954.

[KM97] Kubat M, Matwin S: Addressing the curse of imbalanced training sets: One-sided selection, In: *Proceedings of the Fourteenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, 1997, pp. 179-186.

[Lau01] Laurikkala J: Instance-based data reduction for improved identification of difficult small classes, *Intelligent Data Analysis*. (accepted for publication)

[Lav99] Lavrač N: Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine* 16:1 (1999) 3-23.

[Lie97] Liebowitz J: Worldwide perspectives and trends in expert systems: An analysis based on the three world congresses on expert systems, *AI Magazine* 18:2 (1997) 115-119.

[Lie98] Liebowitz J: *The Handbook of Applied Expert Systems*, CRC Press, Boca Raton, 1998.

[LJ01] Laurikkala J, Juhola M: Hierarchical clustering of female urinary incontinence data having noise and outliers. In: Crespo J, Maojo V, Martin F (eds): *Medical Data Analysis: Proceedings of the 2nd International Symposium (ISMDA 2001)*, Lecture Notes in Computer Science, vol. 2199, Springer, Berlin, 2001. (in press)

[LJK00] Laurikkala J, Juhola M, Kentala E: Informal identification of outliers in medical data. In: Lavrač N, Miksch S, Kavšek B (eds.): *Workshop Notes of the 14th European Conference on Artificial Intelligence (ECAI-2000): The Fifth Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2000)*, Berlin, 2000, pp. 20-24.

[LJP97+] Laurikkala J, Juhola M, Penttinen J, Aukee P: Parameter evaluation of the differential diagnosis of female urinary incontinence for the construction of an expert system. In: Pappas C, Maglaveras N, Scherrer J-R (eds.): *Medical Informatics Europe'97*, Studies in Health Technology and Informatics, vol. 43, IOS Press, Amsterdam, 1997, pp. 671-675.

[LLS00] Lim T-J, Loh W-Y, Shih Y-S: A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine Learning* 40 (2000) 203-228.

[LR87] Little RJA, Rubin DB: *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

[LS95] Langley P, Simon HA: Applications of machine learning and rule induction, *Communications of the ACM* 38:11 (1995) 55-64.

[Mac95] Macartney FJ: Diagnostic logic, In: Phillips CI (ed.): *Logic in Medicine*, BMJ Publishing Group, Plymouth, 1995, pp. 59-99.

[Mic83] Michalski RS: A theory and methodology of inductive learning, In: Michalski RS, Garbonell JG, Mitchell TM (eds.): *Machine Learning: An Artificial Intelligence Approach*, vol. 1, Morgan Kaufmann, Los Altos, 1983, pp. 83-134.

[Mic96] Michalewicz Z: *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, New York, 3rd edn., 1996.

[Mil94] Miller RA: Medical diagnostic decision support systems - past, present, and future: A threaded bibliography and brief commentary, *Journal of the American Medical Informatics Association* 1:1 (1994) 8-27.

[Mit96] Mitchell M: *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, 1996.

[Mit97] Mitchell TM: *Machine Learning*, McGraw-Hill, New York, 1997.

[MK90] Michalski RS, Kodratoff Y: Research in machine learning: Recent progress, classification of methods, and future directions. In: Michalski RS, Kodratoff Y (eds.): *Machine Learning: An Artificial Intelligence Approach*, vol. 3, Morgan Kaufmann, San Mateo, 1990, pp. 3-30.

[Nyk00] Nykänen P: *Decision Support Systems from a Health Informatics Perspective*, Ph.D. Thesis, Department of Computer and Information Sciences, University of Tampere, Finland, 2001. [http://acta.uta.fi/pdf/951-44-4897-9.pdf]

[PEJ96] Pesonen E, Eskelinen M, Juhola M: Comparison of different neural network algorithms in the diagnosis of acute appendicitis, *International Journal of Bio-Medical Computing* 40 (1996) 227-233.

[Pet97] Pett MA: *Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions*, SAGE Publications, Thousand Oaks, 1997.

[PFK98] Provost F, Fawcett T, Kohlavi R: The case against accuracy estimation for comparing induction algorithms. In: Shavlik J (ed.): *Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, 1998, pp. 445-453.

[Qui86] Quinlan JR: Induction of decision trees, *Machine Learning* 1 (1986) 81-106.

[Qui93] Quinlan JR: *C4.5 Programs for Machine Learning*, San Mateo, Morgan Kaufmann, 1993.

[RK88] Riss PA, Koelbl H: Development of an expert system for preoperative assessment of female urinary incontinence, *International Journal of Biomedical Computing* 22 (1988) 217-223.

[RWL94] Rumelhart DE, Widrow B, Lehr MA: The basic ideas in neural networks, *Communications of the ACM* 37:3 (1994) 87-92.

[Sal99] Salzberg SL: On comparing classifiers: A critique of current research and methods, *Data Mining and Knowledge Discovery* 1 (1999) 1-22.

[SDA75+] Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN: Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system, *Computers and Biomedical Research* 8 (1975) 303-320.

[Sha96] Sharma S: *Applied Multivariate Techniques*, Wiley, New York, 1996.

[SN98] Saitta L, Neri F: Learning in the "real world", *Machine Learning* 30 (1998) 133-163.

[SO98] Schafer JL, Olsen MK: Multiple imputation for multivariate missing-data problems: A data analyst's perspective, *Multivariate Behavioral Research* 33:4 (1998) 545-571.

[Swi96] Swingler K: *Applying Neural Networks: A Practical Guide*, Academic Press, London, 1996.

[Tsu98] Tsumoto S: Automated knowledge acquisition from clinical databases based on rough sets and attribute-oriented generalization. In: Chute CG (ed.): *Proceedings the 1998 AMIA Annual Symposium*, Hanley & Belfus, Philadelphia, 1998, pp. 548-552

[Tur93] Turban E: *Decision Support and Expert Systems: Management Support Systems*, Macmillan, New York, 3rd edn., 1993.

[Uri92] Urinary Incontinence Guideline Panel: *Urinary Incontinence in Adults*, U.S. Department of Health and Human Services, 1992, pp. 57-58.

[Wat86] Waterman DA: *A Guide to Expert Systems*, Reading, Addison-Wesley, 1986.

[vBM97] van Bemmel JH, Musen MA (eds.): *Handbook of Medical Informatics*, Springer, Heidelberg, 1997.

[Wei85] Weisberg S: *Applied Linear Regression*, Wiley, New York, 2nd edn., 1985.

[WM97] Wilson DR, Martinez TR: Improved heterogeneous distance functions, *Journal of Artificial Intelligence Research* 6:1 (1997) 1-34.

[WM00] Wilson DR, Martinez TR: Reduction techniques for instance-based learning algorithms, *Machine Learning* 38 (2000) 257-286.

[WWZ99] Webb GI, Wells J, Zheng Z: An experimental evaluation of integrating machine learning with knowledge acquisition, *Machine Learning* 35 (1999) 5-23.

[YZ96] Yuan Y, Zhuang H: A genetic algorithm for generating fuzzy classification rules, *Fuzzy Sets and Systems* 84 (1996) 1-19.

[ZLK99] Zupan B, Lavrač N, Keravnou ET: Special Issue 'Data mining techniques and applications in medicine', *Artificial Intelligence in Medicine* 16:1 (1999) 1-120.

Paper I

Laurikkala J, Juhola M, Lammi S, Penttinen J, Aukee P: Analysis of the imputed female urinary incontinence data for the evaluation of expert system parameters, *Computers in Biology and Medicine* 31 (2001) 239-257.

Paper II

Laurikkala J, Juhola M: A genetic-based machine learning system to discover the diagnostic rules for female urinary incontinence, *Computer Methods and Programs in Biomedicine* 55 (1998) 217-228.

Paper III

Laurikkala J, Juhola M, Lammi S, Viikki K: Comparison of genetic algorithms and other classification methods in the diagnosis of female urinary incontinence, *Methods of Information in Medicine* 38 (1999) 125-131.

Paper IV

Laurikkala J, Juhola M: Nearest neighbour classification with heterogeneous proximity functions. In: Hasman A, Blobel B, Dudeck J, Engelbrecht R, Gell G, Prokosch H-U (eds.): *Medical Infobahn for Europe: Proceedings of MIE2000 and GMDS2000*, Studies in Health Technology and Informatics, vol. 77, IOS Press, Amsterdam, 2000, pp. 753-757.

Paper V

Laurikkala J: Improving identification of difficult small classes by balancing class distribution. In: Quaglini S, Barahona P, Andreassen S (eds.): *Artificial Intelligence in Medicine: Eight European Conference on Artificial Intelligence in Medicine in Europe*, Lecture Notes in Artificial Intelligence, vol. 2101, Springer, Berlin, 2001, pp. 63-66.

Paper VI

Laurikkala J: An Internet-based multi-expert system for the differential diagnosis of female urinary incontinence, *Medical Informatics & The Internet in Medicine*. (submitted)