58

Data Mining in Medicine

Nada Lavrač[1] and Blaž Zupan[2]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia,
Nova Gorica Polytechnic, Vipavska 13, 5000 Nova Gorica, Slovenia
[2] Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
Department of Molecular and Human Genetics, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA

Summary. Extensive amounts of data stored in medical databases require the development of specialized tools for accessing the data, data analysis, knowledge discovery, and effective use of stored knowledge and data. This chapter focuses on Data Mining methods and tools for knowledge discovery. The chapter sketches the selected Data Mining techniques, and illustrates their applicability to medical diagnostic and prognostic problems.

Key words: Data Mining in Medicine, Inductive Logic Programming, Decision Trees, Rule Induction, Case-based Reasoning, Instance-based Learning, Supervised Learning, Neural Networks

58.1 Introduction

Extensive amounts of knowledge and data stored in medical databases require the development of specialized tools for accessing the data, data analysis, knowledge discovery, and effective use of stored knowledge and data, since the increase in data volume causes difficulties in extracting useful information for decision support. The traditional manual data analysis has become insufficient, and methods for efficient computer-based analysis indispensable, such as the technologies developed in the area of *Data Mining* and *knowledge discovery in databases* (Frawley, 1991).

Knowledge discovery in databases is frequently defined as a *process* (Fayyad, 1996) consisting of the following steps: understanding the domain, forming the data set and cleaning the data, extracting of regularities hidden in the data thus formulating knowledge in the form of patterns or models (this step is referred to as *Data Mining* (DM)), postprocessing of discovered knowledge, and exploiting the results.

Important issues that arise from the rapidly emerging globality of data and information are:

O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed.,
DOI 10.1007/978-0-387-09823-4_58, © Springer Science+Business Media, LLC 2010

- the provision of standards in terminology, vocabularies and formats to support multilinguality and sharing of data,
- standards for the abstraction and visualization of data,
- standards for interfaces between different sources of data,
- integration of heterogeneous types of data, including images and signals, and
- reusability of data, knowledge, and tools.

Many environments still lack standards, which hinders the use of data analysis tools on large global data sets, limiting their application to data sets collected for specific diagnostic, screening, prognostic, monitoring, therapy support or other patient management purposes. The emerging standards that relate to Data Mining are CRISP-DM and PMML. CRISP-DM is a Data Mining process standard that was crafted by Cross-Industry Standard Process for Data Mining Interest Group (www.crisp-dm.org). PMML (Predictive Data Mining Markup Language, www.dmg.org), on the other hand, is a standard that defines how to use XML markup language to store predictive Data Mining models, such as classification trees and classification rule sets.

Modern hospitals are well equipped with monitoring and other data collection devices which provide relatively inexpensive means to collect and store the data in inter- and intra-hospital information systems. Large collections of medical data are a valuable resource from which potentially new and useful knowledge can be discovered through Data Mining. Data Mining is increasingly popular as it is aimed at gaining an insight into the relationships and patterns hidden in the data.

Patient records collected for diagnosis and prognosis typically encompass values of anamnestic, clinical and laboratory parameters, as well as results of particular investigations, specific to the given task. Such data sets are characterized by their incompleteness (missing parameter values), incorrectness (systematic or random noise in the data), sparseness (few and/or non-representable patient records available), and inexactness (inappropriate selection of parameters for the given task). The development of Data Mining tools for medical diagnosis and prediction was frequently motivated by the requirements for dealing with these characteristics of medical data sets (Bratko and Kononenko, 1987, Cestnik *et al*., 1987).

Data sets collected in monitoring (either acute monitoring of a particular patient in an intensive care unit, or discrete monitoring over long periods of time in the case of patients with chronic diseases) have additional characteristics: they involve the measurements of a set of parameters at different times, requesting the temporal component to be taken into account in data analysis. These data characteristics need to be considered in the design of analysis tools for prediction, intelligent alarming and therapy support.

In medicine, Data Mining can be used for solving descriptive and predictive Data Mining tasks. *Descriptive* Data Mining tasks are concerned with finding interesting patterns in the data, as well as interesting clusters and subgroups of data, where typical methods include association rule learning, and (hierarchical or k-means) clustering, respectively. In contrast, *predictive* Data Mining starts from the entire data set and aims at inducing a predictive model that holds on the data and can be used for prediction or classification of yet unseen instances. Learning in the predictive Data Mining setting requires labelled data items. Class labels can be either categorical or continuous; accordingly, predictive tasks concern building classification models or regression models, respectively.

Data Mining in medicine is most often used for building classification models, these being used for either diagnosis, prognosis or treatment planning. Predictive Data Mining, which is the focus of this chapter, is concerned with the analysis of classificatory properties of data tables. Data represented in the tables may be collected from measurements or acquired from experts. Rows in the table usually correspond to individuals (training examples) to be analyzed in terms of their properties (attributes) and the class (concept) to which they belong. In a

medical setting, a concept of interest can be a disease or a medical outcome. Supervised learning assumes that training examples are classified whereas unsupervised learning concerns the analysis of unclassified examples.

This chapter is organized as follows. Section 58.2 presents a selection of symbolic classification methods. Section 58.3 complements it by outlining selected subsymbolic classification methods. Finally, Section 58.4 concludes with a brief outline of other methods for supporting medical knowledge discovery.

58.2 Symbolic Classification Methods

In medical data analysis it is very important that the results of data mining can be communicated to humans in an understandable way. In this respect, the analysis tools have to deliver transparent results and preferably facilitate human intervention in the analysis process. A good example of such methods are symbolic machine learning algorithms that, as a result of data analysis, aim to derive a symbolic model (e.g., a decision tree or a set of rules) of preferably low complexity but high transparency and accuracy.

58.2.1 Rule Induction

If-then Rules

Given a set of classified examples, a rule induction system constructs a set of rules. An if-then rule has the form:

```
IF Condition THEN Conclusion.
```

The condition of a rule contains one or more attribute tests of the form $A_i = v_{i_k}$ for discrete attributes, and $A_i < v$ or $A_i > v$ for continuous attributes. The condition of a rule is a conjunction of attribute tests (or a disjunction of conjunctions of attribute tests). The conclusion has the form $C = c_i$, assigning a particular value c_i to class C. An example is *covered* by a rule if the attribute values of the example satisfy the condition in the antecedent of the rule.

An example rule below, induced in the domain of early diagnosis of rheumatic diseases (Lavrač *et al*., 1993, Džeroski and Lavrač, 1996), assigns the diagnosis crystal-induced synovitis to male patients older than 46 who have more than three painful joints and psoriasis as a skin manifestation.

```
IF    Sex = male
  AND Age > 46
  AND Number_of_painful_joints > 3
  AND Skin_manifestations = psoriasis
THEN  Diagnosis = crystal_induced_synovitis
```

If-then rule induction, studied already in the eighties (Michalski, 1986), resulted in a series of AQ algorithms, including the AQ15 system which was applied also to the analysis of medical data (Michalski et al. 1986).

Here we describe the rule induction system CN2 (Clark and Niblett, 1989, Clark and Boswell, 1991) which is among the best known if-then rule learners capable of handling imperfect/noisy data. Like the AQ algorithms, CN2 also uses the covering approach to construct a set of rules for each possible class c_i in turn: when rules for class c_i are being constructed, examples of this class are treated as positive, and all other examples as negative. The covering approach works as follows: CN2 constructs a rule that correctly classifies some positive examples, removes the positive examples covered by the rule from the training set and repeats the process until no more positive examples remain uncovered. To construct a single rule that classifies examples into class c_i, CN2 starts with a rule with an empty condition (IF part) and the selected class c_i as the conclusion (THEN part). The antecedent of this rule is satisfied by all examples in the training set, and not only those of the selected class. CN2 then progressively refines the antecedent by adding conditions to it, until only examples of class c_i satisfy the antecedent. To allow for the handling imperfect data, CN2 may construct a set of rules which is imprecise, i.e., does not classify all examples in the training set correctly.

Consider a partially built rule. The conclusion part is fixed to c_i and there are some (possibly none) conditions in the IF part. The examples covered by this rule form the current training set. For discrete attributes, all conditions of the form $A_i = v_{i_k}$, where v_{i_k} is a possible value for A_i, are considered for inclusion in the condition part. For continuous attributes, all conditions of the form $A_i \leq \frac{v_{i_k}+v_{i_{k+1}}}{2}$ and $A_i > \frac{v_{i_k}+v_{i_{k+1}}}{2}$ are considered, where v_{i_k} and $v_{i_{k+1}}$ are two consecutive values of attribute A_i that actually appear in the current training set. For example, if the values 4.0, 1.0, and 2.0 for attribute A_i appear in the current training set, the conditions $A_i \leq 1.5$, $A_i > 1.5$, $A_i \leq 3.0$, and $A_i > 3.0$ will be considered.

Note that both the structure (set of attributes to be included) and the parameters (values of the attributes for discrete ones and boundaries for the continuous ones) of the rule are determined by CN2. Which condition will be included in the partially built rule depends on the number of examples of each class covered by the refined rule and the heuristic estimate of the quality of the rule.

The heuristic estimates used in rule induction are mainly designed to estimate the performance of the rule on unseen examples in terms of classification accuracy. This is in accordance with the task of achieving high classification accuracy on unseen cases. Suppose a rule covers p positive and n negative examples of class c_j. Its accuracy an be estimated by the relative frequency of positive examples of class c_j covered, computed as $p/(p+n)$. This heuristic, used in early rule induction algorithms, prefers rules which cover examples of only one class. The problem with this metric is that it tends to select very specific rules supported by few examples. In the extreme case, a maximally specific rule will cover one example and hence have an unbeatable score using the metrics of apparent accuracy (scoring 100% accuracy). Apparent accuracy on the training data, however, does not necessarily reflect true predictive accuracy, i.e., accuracy on new test data. It has been shown (Holte *et al.*, 1989) that rules supported by few examples have very high error rates on new test instances.

The problem lies in the estimation of the probabilities involved, i.e., the estimate of the probability that a new instance is correctly classified by a given rule. If we use relative frequency, the estimate is only good if the rule covers many examples. In practice, however, not enough examples are available to estimate these probabilities reliably at each step. Therefore, probability estimates that are more reliable when few examples are given should be used, such as the Laplace estimate which, in two-class problems, estimates the accuracy as $(p+1)/(p+n+2)$ (Niblett and Bratko, 1986). This is the search heuristic used in CN2. The m-estimate (Cestnik, 1990) is a further upgrade of the Laplace estimate, taking also into account the prior distribution of classes.

Rule induction can be used for early diagnosis of rheumatic diseases (Lavrač *et al*., 1993, Džeroski and Lavrač, 1996), for the evaluation of EDSS in multiple sclerosis (Gaspari *et al*., 2001) and in numerous other medical domains.

Rough Sets

If-then rules can be also induced using the theory of *rough sets* (Pawlak, 1981, Pawlak, 1991). Rough sets (RS) are concerned with the analysis of classificatory properties of data aimed at approximations of concepts. RS can be used both for supervised and unsupervised learning.

Let us introduce the main concepts of the rough set theory. Let U denote a non-empty finite set of *objects* called the *universe* and A a non-empty finite set of *attributes*. Each object $x \in U$ is assumed to be described by a subset of attributes B, $B \subseteq A$. The basic concept of RS is an *indiscernibility* relation. Two objects x and y are indiscernible on the basis of the available attribute subset B if they have the same values of attributes B. It is usually assumed that this relation is reflexive, symmetric and transitive. The set of objects indiscernible from x using attributes B forms an equivalence class and is denoted by $[x]_B$. There are extensions of RS theory that do not require transitivity to hold.

Let $X \subseteq U$, and let $Ind_B(X)$ denote a set of equivalence classes of examples that are indiscernible, i.e., a set of subsets of examples that cannot be distinguished on the basis of attributes in B. The subset of attributes B is sufficient for classification if for every $[x]_B \in Ind_B(X)$ all the examples in $[x]_B$ belong to the same decision class. In this case crisp definitions of classes can be induced; otherwise, only 'rough' concept definitions can be induced since some examples can not be decisively classified.

The goal of RS analysis is to induce approximations of concepts c_i. Let X consist of training examples of class c_i. X may be approximated using only the information contained in B by constructing the *B-lower* and *B-upper approximations of* X, denoted $\underline{B}X$ and $\overline{B}X$ respectively, where $\underline{B}X = \{x \mid x \in X,\ [x]_B \subseteq X\}$ and $\overline{B}X = \{x \mid x \in U,\ [x]_B \cap X \neq \emptyset\}$. On the basis of knowledge in B the objects in $\underline{B}X$ can be classified with certainty as members of X, while the objects in $\overline{B}X$ can be only classified as possible members of X. The set $BN_B(X) = \overline{B}X - \underline{B}X$ is called the *B-boundary region of* X thus consisting of those objects that on the basis of knowledge in B cannot be unambiguously classified into X or its complement. The set $U - \overline{B}X$ is called the *B-outside region of* X and consists of those objects which can be with certainty classified as not belonging to X. A set is said to be *rough* (respectively *crisp*) if the boundary region is non-empty (respectively empty). The boundary region consists of examples that are indiscernible from some examples in X and therefore can not be decisively classified into c_i; this region consists of the union of equivalence classes each of which contains some examples from X and some examples not in X.

The main task of RS analysis is to find minimal subsets of attributes that preserve the indiscernibility relation. This is called the *reduct* computation. Note that there are usually many reducts. Several types of reducts exist. Decision rules are generated from reducts by reading off the values of the attributes in each reduct. The main challenge in inducing rules lies in determining which attributes should be included in the condition of the rule. Rules induced from the (standard) reducts will usually result in large sets of rules and are likely to overfit the data. Instead of standard reducts, attribute sets that "almost" preserve the indiscernibility relation are generated. Good results have been achieved with *dynamic reducts* (Skowron, 1995) that use a combination of reduct computation and statistical resampling. Many RS approaches to discretization, feature selection, symbolic attribute grouping, have also been designed (Polkowski and Skowron, 1998a, Polkowski and Skowron, 1998b). There exist also several software tools for RS, such as the Rosetta system (Rumelhart, 1986).

The list of applications of RS in medicine is significant. It includes extracting diagnostic rules, image analysis and classification of histological pictures, modelling set residuals, EEG signal analysis, etc (Averbuch *et al*., 2004, Rokach *et al*., 2004). Examples of RS analysis in medicine include (Grzymala-Busse, 1998, Komorowski and Øhrn, 1998, Tsumoto, 1998). For references that include medical applications, see (Polkowski and Skowron, 1998a, Polkowski and Skowron, 1998b, Lin and Cercone, 1997).

Ripple Down Rules

The knowledge representation of the form of ripple down rules allows incremental learning by including exceptions to the current rule set. Ripple down rules (RDR) (Compton and Jansen, 1988, Compton *et al*., 1989) have the following form:

```
IF Conditions THEN Conclusion BECAUSE Case EXCEPT
    IF ...
ELSE IF ...
```

For the domain of lens prescription (Cendrowka, 1987) an example RDR (Sammut, 1998) is shown below.

```
IF true THEN no_lenses BECAUSE case0
    EXCEPT
        IF astigmatism = not_astigmatic and
           tear_production = normal
        THEN
           soft_lenses BECAUSE case2
ELSE
    IF prescription = myope and
       tear_production = normal
    THEN
       hard_lenses BECAUSE case4
```

The contact lenses RDR is interpreted as follows: The default rule is that a person does not use lenses, stored in the rule base together with a 'dummy' case0. No update of the system is needed after entering the data on the first patient who needs no lenses. But the second patient (case2) needs soft lenses and the rule is updated according to the conditions that hold for case2. Case3 is again a patient who does not need lenses, but the rule needs to be updated w.r.t. the conditions of the fourth patient (case4) who needs hard lenses.

The above example illustrates also the incremental learning of ripple down rules in which EXCEPT IF THEN and ELSE IF THEN statements are added to the RDRs to make them consistent with the current database of patients.

If the RDR from example above were rewritten as an IF-THEN-ELSE statement it would look as follows:

```
IF true THEN
    IF astigmatism = not_astigmatic and
       tear_production = normal
    THEN
       soft_lenses ELSE no_lenses
ELSE
    IF prescription = myope and
       tear_production = normal
    THEN
       hard_lenses
```

There were many successful medical applications of the RDR approach, including the system PEIRS (Edwards *et al*., 1993) which is an RDR reconstruction of the hand-built GARVAN expert system knowledge base on thyroid function tests (Horn *et al*., 1985).

58.2.2 Learning of Classification and Regression Trees

Systems for Top-Down Induction of Decision Trees (Quinlan, 1986) generate a decision tree from a given set of examples. Each of the interior nodes of the tree is labelled by an attribute, while branches that lead from the node are labelled by the values of the attribute.

The tree construction process is heuristically guided by choosing the 'most informative' attribute at each step, aimed at minimizing the expected number of tests needed for classification. Let E be the current (initially entire) set of training examples, and $c_1, \ldots, c_N$ the decision classes. A decision tree is constructed by repeatedly calling a tree construction algorithm in each generated node of the tree. Tree construction stops when all examples in a node are of the same class (or if some other stopping criterion is satisfied). This node, called a leaf, is labelled by class value. Otherwise the 'most informative' attribute, say A_i, is selected as the root of the (sub)tree, and the current training set E is split into subsets E_i according to the values of the most informative attribute. Recursively, a subtree T_i is built for each E_i.

Ideally, each leaf is labelled by exactly one class value. However, leaves can also be empty, if there are no training examples having attribute values that would lead to a leaf, or can be labelled by more than one class value (if there are training examples with same attribute values and different class values).

One of the most important features is tree pruning, used as a mechanism for handling noisy data (Quinlan, 1993). Tree pruning is aimed at producing trees which do not overfit possibly erroneous data. In tree pruning, the unreliable parts of a tree are eliminated in order to increase the classification accuracy of the tree on unseen instances.

An early decision tree learner, ASSISTANT (Cestnik *et al*., 1987), that was developed specifically to deal with the particular characteristics of medical data sets, supports the handling of incompletely specified training examples (missing attribute values), binarization of continuous attributes, binary construction of decision trees, pruning of unreliable parts of the tree and plausible classification based on the 'naive' Bayesian principle to calculate the classification in the leaves for which no evidence is available. An example decision tree that can be used to predict outcome of patients after severe head injury (Pilih, 1997) is shown in Figure 58.1. The two attributes in the nodes of the tree are CT score (number of abnormalities

detected by Computer axial Tomography) and GCS (evaluation of coma according to the Glasgow Coma Scale).

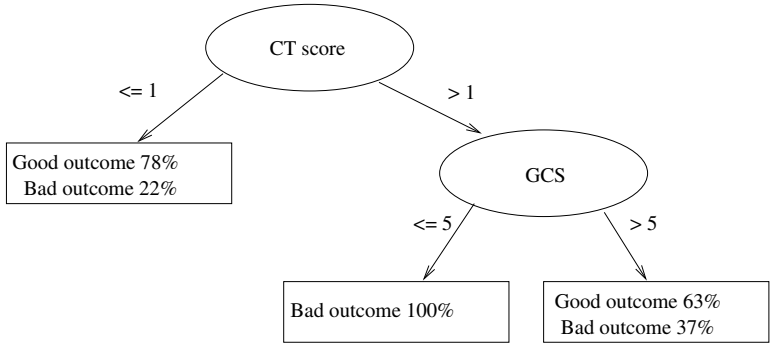

Fig. 58.1. Decision tree for outcome prediction after severe head injury. In the leaves, the percentages indicate the probabilities of class assignment.

Implementations of the ASSISTANT algorithm include ASSISTANT-R and ASSISTANT-R2 (Kononenko and Šimec, 1995). Instead of the standardly used informativity search heuristic, ASSISTANT-R employs ReliefF as a heuristic for attribute selection (Kononenko, 1994, Kira and Rendell, 1992b). This heuristic is an extension of RELIEF (Kira and Rendell, 1992a, Kira and Rendell, 1992b) which is a non-myopic heuristic measure that is able to estimate the quality of attributes even if there are strong conditional dependencies between attributes. In addition, wherever appropriate, instead of the relative frequency, ASSISTANT-R uses the *m*-estimate of probabilities (Cestnik, 1990).

The best known decision tree learner is C4.5 (Quinlan, 1993) (See5 and J48 are its more recent upgrades) which is widely used and has been incorporated into commercial Data Mining tools as well as in the publicly available WEKA Data Mining toolbox (Witten and Frank, 1999). The system is reliable, efficient and capable of dealing with large sets of training examples.

Learning of regression trees is similar to decision tree learning: it also uses a top-down greedy approach to tree construction. The main difference is that decision tree construction involves the classification into a finite set of discrete classes whereas in regression tree learning the decision variable is continuous and the leaves of the tree either consist of a prediction into a numeric value or a linear combination of variables (attributes). An early learning system CART (Breiman *et al*., 1984) featured both classification and regression tree learning.

There are many applications of decision trees for analysis of medical data sets. For instance, CART has been applied to the problem of mining a diabetic data warehouse composed of a complex relational database with time series and sequencing information (Breault and Goodall, 2002). Decision tree learning has been applied to the diagnosis of sport injuries (Zelic *et al.*, 1997), patient recovery prediction after traumatic brain injury (Andrews *et al*., 2002), prediction of recurrent falling in community-dwelling older persons (Stel *et al*., 2003), and numerous other medical domains.

58.2.3 Inductive Logic Programming

Inductive logic programming (ILP) systems learn relational concept descriptions from relational data. Well known ILP systems include FOIL (Quinlan, 1990), Progol (Muggleton, 1995) and Claudien (De Raedt and Dehaspe, 1997). LINUS is an ILP environment (Lavrač and Džeroski, 1994), enabling the transformation of relational learning problems into the form appropriate for standard attribute-value learners, while in general ILP systems learn relational descriptions without such a transformation to propositional learning.

In ILP, induced rules typically have the form of Prolog clauses. The output of an ILP system is illustrated by a rule of ocular fundus image classification for glaucoma diagnosis, induced by an ILP system GKS (Mizoguchi *et al*., 1997) specially designed to deal with low-level measurement data including images.

```
class(Image, Segment, undermining) :-
   clockwise(Segment, Adjacent, 1),
   class_confirmed(Image, Adjacent, undermining).
```

Compared to rules induced by a rule learning algorithm of the form IF *Condition* THEN *Conclusion*, Prolog rules have the form *Conclusion* :- *Condition*. For example, the rule for glaucoma diagnosis means that Segment of Image is classified as undermining (i.e., not normal) if the conditions of the right-hand side of the clause are fulfilled. Notice that the conditions consist of a conjunction of predicate clockwise/3 defined in the background knowledge, and predicate class_confirmed/3, added to the background knowledge in one of the previous iterative runs of the GKS algorithm. This shows one of the features of ILP learning, namely that learning can be done in several cycles of the learning algorithm in which definitions of new background knowledge predicates are learned and used in the subsequent runs of the learner; this may improve the performance of the learner.

ILP has been successfully applied to carcinogenesis prediction in the predictive toxicology evaluation challenge (Srinivasan *et al*., 1997) and to the recognition of arrhythmia from electrocardiograms (Carrault *et al*., 2003).

58.2.4 Discovery of Concept Hierarchies and Constructive Induction

The data can be decomposed into equivalent but smaller, more manageable and potentially easier to comprehend data sets. A method that uses such an approach is called *function decomposition* (Zupan and Bohanec, 1998). Besides the discovery of appropriate data sets, function decomposition arranges them into a concept hierarchy. Function decomposition views classification data (example set) with attributes $X = \{x_1, \ldots, x_n\}$ and an output concept (class) y defined as a partially specified function $y = F(X)$. The core of the method is a single step decomposition of F into $y = G(A, c)$ and $c = H(B)$, where A and B are proper subsets of input attributes such that $A \cup B = X$. Single step decomposition constructs the example sets that partially specify new functions G and H. Functions G and H are determined in the decomposition process and are not predefined in any way. Their joint complexity (determined by some complexity measure) should be lower than the complexity of F. Obviously, there are many candidates for partitioning X into A and B; the decomposition chooses the partition that yields functions G and H of lowest complexity. In this way, single step decomposition also discovers a new intermediate concept $c = H(B)$. Since the decomposition can be applied recursively

on H and G, the result in general is a hierarchy of concepts. For each concept in the hierarchy, there is a corresponding function (such as $H(B)$) that determines the dependency of that concept on its immediate descendants in the hierarchy.

In terms of data analysis, the benefits of function decompositions are:

- Discovery of new data sets that use fewer attributes than the original one and include fewer instances as well. Because of lower complexity, such data sets may then be easier to analyze.
- Each data set represents some concept. Function decomposition organizes discovered concepts in a hierarchy, which may itself be interpretable and can help to gain insight into the data relationships and underlying attribute groups.

Consider for example a concept hierarchy in Figure 58.2 that was discovered for a data set that describes a nerve fiber conduction-block (Zupan *et al.*, 1997). The original data set used 2543 instances of six attributes (aff, nl, k-conc, na-conc, scm, leak) and a single class variable (block) determining nerve fiber conducts or not. Function decomposition found three intermediate concepts, c1, c2, and c3. When interpreted by the domain expert, it was found that the discovered intermediate concepts are physiologically meaningful and constitute useful intermediate biophysical properties. Intermediate concept c1, for example, couples the concentration of ion channels (na-conc and k-conc) and ion leakage (leak) that are all the axonal properties and together influence the combined current source/sink capacity of the axon which is the driving force for all propagated action potentials. Moreover, new concepts use fewer attributes and instances: c1, c2, c3, and the output concept block described 125, 25, 184, and 65 instances, respectively.

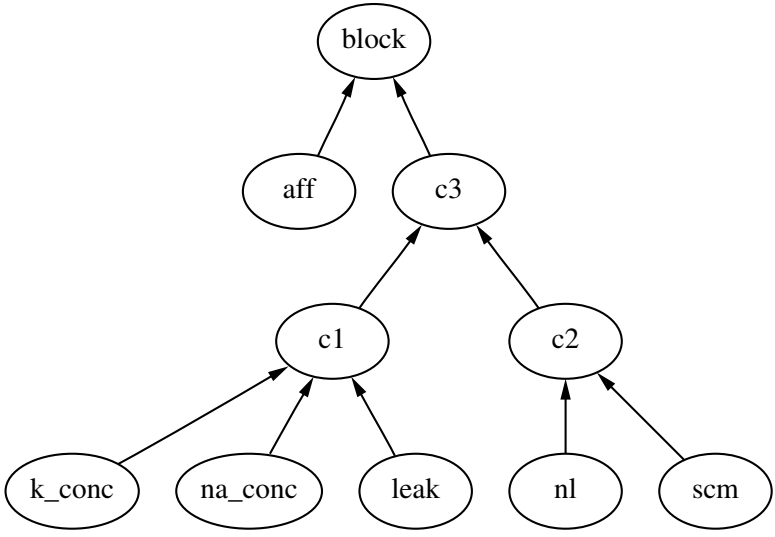

Fig. 58.2. Discovered concept hierarchy for the conduction-block domain.

Intermediate concepts discovered by decomposition can also be regarded as new features that can, for example, be added to the original example set, which can then be examined by

some other data analysis method. Feature discovery and constructive induction, first investigated in (Michalski, 1986), are defined as an ability of the system to derive and use new attributes in the process of learning. Besides pure performance benefits in terms of classification accuracy, constructive induction is useful for data analysis as it may help to induce simpler and more comprehensible models and to identify interesting inter-attribute relationships. New attributes may be constructed based on available background knowledge of the domain: an example of how this facilitated learning of more accurate and comprehensible rules in the domain of early diagnosis of rheumatic diseases is given in (Džeroski and Lavrač, 1996). Function decomposition, on the other hand, may help to discover attributes from classified instances alone. For the same rheumatic domain, this is illustrated in (Zupan and Džeroski, 1998). Although such discovery may be carried out automatically, the benefits of the involvement of experts in new attribute selection are typically significant (Zupan *et al*., 2001).

58.2.5 Case-Based Reasoning

Case-based reasoning (CBR) uses the knowledge of past experience when dealing with new cases (Aamodt and Plaza, 1994, Macura and Macura, 1997). A "case" refers to a problem situation. Although, as in instance-based learning (Aha *et al*., 1991), cases (examples) can be described by a simple attribute-value vector, CBR most often uses a richer, often hierarchical data structure. CBR relies on a database of past cases that has to be designed in the way to facilitate the retrieval of similar cases. CBR is a four stage process:

1. Given a new case to solve, a set of similar cases is retrieved from the database.
2. The retrieved cases are reused in order to obtain a solution for a new case. This may be simply achieved by selecting the most frequent solution used with similar past cases, or, if appropriate background knowledge or a domain model exist, retrieved solutions may be adapted for a new case.
3. The solution for the new case is then checked by the domain expert, and, if not correct, repaired using domain-specific knowledge or expert's input. The specific revision may be saved and used when solving other new cases.
4. The new case, its solution, and any additional information used for this case that may be potentially useful when solving new cases are then integrated in the case database.

CBR offers a variety of tools for data analysis. The similar past cases are not just retrieved, but are also inspected for most relevant features that are similar or different to the case in question. Because of the hierarchical data organization, CBR may incorporate additional explanation mechanisms. The use of symbolic domain knowledge for solution adaptation may further reveal specifics and interesting case's features. When applying CBR to medical data analysis, however, one has to address several non-trivial questions, including the appropriateness of similarity measures used, the actuality of old cases (as the medical knowledge is rapidly changing), how to handle different solutions (treatment actions) by different physicians, etc.

Several CBR systems were used, adapted for, or implemented to support reasoning and data analysis in medicine. Some are described in the special issue of *Artificial Intelligence in Medicine* (Macura and Macura, 1997) and include CBR systems for reasoning in cardiology by Reategui et al., learning of plans and goal states in medical diagnosis by López and Plaza, detection of coronary heart disease from myocardial scintigrams by Haddad et al., and treatment advice in nursing by Yearwood and Wilkinson. Others include a system that uses CBR to assist in the prognosis of breast cancer (Mariuzzi *et al*., 1997), case classification in the domain of ultrasonography and body computed tomography (Kahn and Anderson, 1994), and

a CBR-based expert system that advises on the identification of nursing diagnoses in a new client (Bradburn *et al*., 1993). There is also an application of case-based distance measurements in coronary interventions (Gyöngyösi, 2002).

58.3 Subsymbolic Classification Methods

In medical problem solving it is important that a decision support system is able to explain and justify its decisions. Especially when faced with an unexpected solution of a new problem, the user requires substantial justification and explanation. Hence the interpretability of induced knowledge is an important property of systems that induce solutions from data about past solved cases. Symbolic Data Mining methods have this property since they induce symbolic representations (such as decision trees) from data. On the other hand, subsymbolic Data Mining methods typically lack this property which hinders their use in situations for which explanations are required. Nevertheless, when classification accuracy is the main applicability criterion subsymbolic methods may turn out to be very appropriate since they typically achieve accuracies that are at least as good as those of symbolic classifiers.

58.3.1 Instance-Based Learning

Instance-based learning (IBL) algorithms (Aha *et al*., 1991) use specific instances to perform classification, rather than generalizations induced from examples, such as induced if-then rules. IBL algorithms are also called lazy learning algorithms, as they simply save some or all of the training examples and postpone all the inductive generalization effort until classification time. They assume that similar instances have similar classifications: novel instances are classified according to the classifications of their most similar neighbors.

IBL algorithms are derived from the nearest neighbor pattern classifier (Fix and Hodges, 1957, Cover and Hart, 1968). The nearest neighbor (NN) algorithm is one of the best known classification algorithms; an enormous body of research exists on the subject (Dasarathy, 1990). In essence, the NN algorithm treats attributes as dimensions of an Euclidean space and examples as points in this space. In the training phase, the classified examples are stored without any processing. When classifying a new example, the Euclidean distance between this example and all training examples is calculated and the class of the closest training example is assigned to the new example.

The more general k-NN method takes the k nearest training examples and determines the class of the new example by majority vote. In improved versions of k-NN, the votes of each of the k nearest neighbors are weighted by the respective proximity to the new example (Dudani, 1975). An optimal value of k may be determined automatically from the training set by using leave-one-out cross-validation (Weiss and Kulikowski, 1991). In the k-NN algorithm implementation described in (Wettschereck, 1994), the best k from the range [1,75] was selected in this manner. This implementation also incorporates feature weights determined from the training set. Namely, the contribution of each attribute to the distance may be weighted, in order to avoid problems caused by irrelevant features (Wolpert, 1989).

Let $n = N_{at}$. Given two examples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, the distance between them is calculated as

$$\text{distance}(x, y) = \sqrt{\sum_{i=1}^{n} w_i \cdot \text{difference}(x_i, y_i)^2} \tag{58.1}$$

where w_i is a non-negative weight value assigned to feature (attribute) A_i and the difference between attribute values is defined as follows

$$\text{difference}(x_i, y_i) = \begin{cases} |x_i - y_i| & \text{if } A_i \text{ is continuous} \\ 0 & \text{if } A_i \text{ is discrete and } x_i = y_i \\ 1 & \text{otherwise} \end{cases} \quad (58.2)$$

When classifying a new instance z, k-NN selects the set K of k-nearest neighbors according to the distance defined above. The vote of each of the k nearest neighbors is weighted by its proximity (inverse distance) to the new example. The probability $p(z, c_j, K)$ that instance z belongs to class c_j is estimated as

$$p(z, c_j, K) = \frac{\sum_{x \in K} x_{c_j} / \text{distance}(z, x)}{\sum_{x \in K} 1 / \text{distance}(z, x)} \quad (58.3)$$

where x is one of the k nearest neighbors of z and x_{c_j} is 1 if x belongs to class c_j. Class c_j with largest value of $p(z, c_j, K)$ is assigned to the unseen example z.

Before training (respectively before classification), the continuous features are normalized by subtracting the mean and dividing by the standard deviation so as to ensure that the values output by the difference function are in the range [0,1]. All features have then equal maximum and minimum potential effect on distance computations. However, this bias handicaps k-NN as it allows redundant, irrelevant, interacting or noisy features to have as much effect on distance computation as other features, thus causing k-NN to perform poorly. This observation has motivated the creation of many methods for computing feature weights.

The purpose of a feature weight mechanism is to give low weight to features that provide no information for classification (e.g., very noisy or irrelevant features), and to give high weight to features that provide reliable information. In the k-NN implementation of Wettschereck (Wettschereck, 1994), feature A_i is weighted according to the mutual information (Shannon, 1948) $I(c_j, A_i)$ between class c_j and attribute A_i.

Instance-based learning was applied to the problem of early diagnosis of rheumatic diseases (Džeroski and Lavrač, 1996).

58.3.2 Neural Networks

Artificial neural networks can be used for both supervised and unsupervised learning. For each learning type, we briefly describe the most frequently used approaches.

Supervised Learning

For supervised learning and among different neural network paradigm, feed-forward multilayered neural networks (Rumelhart and McClelland, 1986, Fausett, 1994) are most frequently used for modeling medical data. They are computational structures consisting of a interconnected processing elements (PE) or nodes arranged on a multi-layered hierarchical architecture. In general, a PE computes the weighted sum of its inputs and filters it through some sigmoid function to obtain the output (Figure 58.3.a). Outputs of PEs of one layer serve as inputs to PEs of the next layer (Figure 58.3.b). To obtain the output value for selected instance, its attribute values are stored in input nodes of the network (the network's lowest layer). Next, in each step, the outputs of the higher-level processing elements are computed (hence the name feed-forward), until the result is obtained and stored in PEs at the output layer.

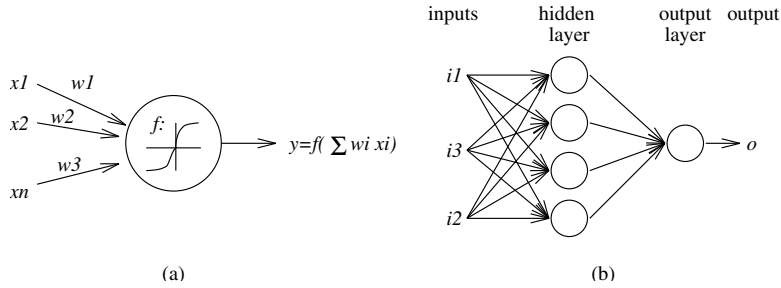

Fig. 58.3. Processing element (a) and an example of the typical structure of the feed-forward multi-layered neural network with four processing elements at hidden layer and one at output layer (b).

A typical architecture of multi-layered neural network comprising an input, a hidden and and output layer of nodes is given in Figure 58.3.b. The number of nodes in the input and output layers is domain-dependent and, respectively, is related to number and type of attributes and a type of classification task. For example, for a two-class classification problem, a neural net may have two output PEs, each modelling the probability of a distinct class, or a single PE, if a problem is coded properly.

Weights that are associated with each node are determined from training instances. The most popular learning algorithm for this is backpropagation (Rumelhart and McClelland, 1986, Fausett, 1994). Backpropagation initially sets the weights to some arbitrary value, and then considering one or several training instances at the time adjusts the weights so that the error (difference between the expected and the obtained value of nodes at the output level) is minimized. Such a training step is repeated until the overall classification error across all of the training instances falls below some specified threshold.

Most often, a single hidden layer is used and the number of nodes has to be either defined by the user or determined through learning. Increasing the number of nodes in a hidden layer allows more modeling flexibility but may cause overfitting of the data. The problem of determining the "right architecture", together with the high complexity of learning, are two of the limitations of feed-forward multi-layered neural networks. Another is the need for proper preparation of the data (Kattan and Beck, 1995): a common recommendation is that all inputs are scaled over the range from 0 to 1, which may require normalization and encoding of input attributes.

For data analysis tasks, however, the most serious limitation is the lack of explanational capabilities: the induced weights together with the network's architecture do not usually have an obvious interpretation and it is usually difficult or even impossible to explain "why" a certain decision was reached. Recently, several approaches for alleviating this limitation have been proposed. A first approach is based on pruning of the connections between nodes to obtain sufficiently accurate, but in terms of architecture significantly less complex, neural networks (Chung and Lee, 1992). A second approach, which is often preceded by the first one to reduce the complexity, is to represent a learned neural network with a set of symbolic rules (Andrews *et al.*, 1995, Craven and Shavlik, 1997, Setiono, 1997, Setiono, 1999).

Despite the above-mentioned limitations, multi-layered neural networks often have equal or superior predictive accuracy when compared to symbolic learners or statistical approaches (Kattan and Beck, 1995, Shawlik *et al.*, 1991). They have been extensively used to model

medical data. Example applications areas include survival analysis (Liestøl *et al*., 1994), clinical medicine (Baxt, 1995), pathology and laboratory medicine (Astion and Wilding, 1992), molecular sequence analysis (Wu, 1997), pneumonia risk assessment (Caruana *et al*., 1995), and prostate cancer survival (Kattan *et al*., 1997). There are fewer applications where rules were extracted from neural networks: an example of such data analysis is finding rules for breast cancer diagnosis (Setiono, 1996).

Different types of neural networks for supervised learning include Hopfield's recurrent networks and neural networks based on adaptive resonance theory mapping (ARTMAP). For the first, an example application is tumor boundary detection (Zhu and Yan, 1997). Example studies of application of ARTMAP in medicine include classification of cardiac arrhythmias (Ham and Han, 1996) and treatment selection for schizophrenic and unipolar depressed in-patients (Modai *et al*., 1996). Learned ARTMAP networks can also be used to extract symbolic rules (Carpenter and Tan, 1993, Downs *et al*., 1996). There are numerous medical applications of neural networks, including brain volumes characterization (Bona *et al*., 2003).

Unsupervised Learning

For unsupervised learning — learning which is presented with unclassified instances and aims at identifying groups of instances with similar attribute values — the most frequently used neural network approach is that of Kohonen's self organizing maps (SOM) (Kohonen, 1988). Typically, SOM consist of a single layer of output nodes. An output node is fully connected with nodes at the input layer. Each such link has an associated weight. There are no explicit connections between nodes of the output layer.

The learning algorithm initially sets the weights to some arbitrary value. At each learning step, an instance is presented to the network, and a winning output node is chosen based on instance's attribute values and node's present weights. The weights of the winning node and of the topologically neighboring nodes are then updated according to their present weights and instance's attribute values. The learning results in the internal organization of SOM such that when two similar instances are presented, they yield a similar "pattern" of networks output node values. Hence, data analysis based on SOM may be additionally supported by proper visualization methods that show how the patterns of output nodes depend on input data (Kohonen, 1988). As such, SOM may not only be used to identify similar instances, but can, for example, also help to detect and analyze time changes of input data. Example applications of SOM include analysis of ophthalmic field data (Henson *et al*., 1997), classification of lung sounds (Malmberg *et al*., 1996), clinical gait analysis (Koehle *et al*., 1997), analysis of molecular similarity (Barlow, 1995), and analysis of a breast cancer database (Markey *et al*., 2002).

58.3.3 Bayesian Classifier

The Bayesian classifier uses the naive Bayesian formula to calculate the probability of each class c_j given the values v_{i_k} of all the attributes for a given instance to be classified (Kononenko, 1993, 1). For simplicity, let $(v_1, \ldots, v_n)$ denote the n-tuple of values of example e_k to be classified. Assuming the conditional independence of the attributes given the class, i.e., assuming $p(v_1..v_n|c_j) = \prod_i p(v_i|c_j)$, then $p(c_j|v_1..v_n)$ is calculated as follows:

$$p(c_j|v_1..v_n) = \frac{p(c_j.v_1..v_n)}{p(v_1..v_n)} = \frac{p(v_1..v_n|c_j) \cdot p(c_j)}{p(v_1..v_n)} = \tag{58.4}$$

$$\frac{\prod_i p(v_i|c_j)\cdot p(c_j)}{p(v_1..v_n)} = \frac{p(c_j)}{p(v_1..v_n)}\prod_i \frac{p(c_j|v_i)\cdot p(v_i)}{p(c_j)} =$$

$$p(c_j)\frac{\prod_i p(v_i)}{p(v_1..v_n)}\prod_i \frac{p(c_j|v_i)}{p(c_j)}$$

A new instance will be classified into the class with maximal probability.

In the above equation, $\frac{\prod_i p(v_i)}{p(v_1..v_n)}$ is a normalizing factor, independent of the class; it can therefore be ignored when comparing values of $p(c_j|v_1..v_n)$ for different classes c_j. Hence, $p(c_j|v_1..v_n)$ is proportional to:

$$p(c_j)\prod_i \frac{p(c_j|v_i)}{p(c_j)} \tag{58.5}$$

Different probability estimates can be used for computing the probabilities, i.e., the relative frequency, the Laplace estimate (Niblett and Bratko, 1986), and the *m*-estimate (Cestnik, 1990, Kononenko, 1993, 1).

Continuous attributes have to be pre-discretized in order to be used by the naive Bayesian classifier. The task of discretization is the selection of a set of boundary values that split the range of a continuous attribute into a number of intervals which are then considered as discrete values of the attribute. Discretization can be done manually by the domain expert or by applying a discretization algorithm (Richeldi and Rossotto, 1995).

The problem of (strict) discretization is that minor changes in the values of continuous attributes (or, equivalently, minor changes in boundaries) may have a drastic effect on the probability distribution and therefore on the classification. Fuzzy discretization may be used to overcome this problem by considering the values of the continuous attribute (or, equivalently, the boundaries of intervals) as fuzzy values instead of point values (Kononenko, 1993). The effect of fuzzy discretization is that the probability distribution is smoother and the estimation of probabilities more reliable, which in turn results in more reliable classification.

Bayesian computation can also be used to support decisions in different stages of a diagnostic process (McSherry, 1997) in which doctors use *hypothetico-deductive reasoning* for gathering evidence which may help to confirm a diagnostic hypothesis, eliminate an alternative hypothesis, or discriminate between two alternative hypotheses. In particular, Bayesian computation can help in identifying and selecting the most useful tests, aimed at confirming the target hypothesis, eliminating the likeliest alternative hypothesis, increase the probability of the target hypothesis, decrease the probability of the likeliest alternative hypothesis or increase the probability of the target hypothesis relative to the likeliest alternative hypothesis. Bayesian classification has been applied to different medical domains, including the diagnosis of sport injuries (Zelic *et al.*, 1997).

58.4 Other Methods Supporting Medical Knowledge Discovery

There is a variety of other methods and tools that can support medical data analysis and can be used separately or in combination with the classification methods introduced above. We here mention only several most frequently used techniques.

The problem of discovering *association rules* has recently received much attention in the Data Mining community. The problem of inducing association rules (Agrawal *et al*., 1996) is defined as follows: Given a set of transactions, where each transaction is a set of items (i.e., literals of the form $Attribute = value$), an *association rule* is an expression of the form $X \rightarrow Y$ where X and Y are sets of items. The intuitive meaning of such a rule is that transactions in

a database which contain X tend to contain Y. Consider a sample association rule: "80% of patients with pneumonia also have high fever. 10% of all transactions contain both of these items." Here 80% is called *confidence* of the rule, and 10% support of the rule. Confidence of the rule is calculated as the ratio of the number of records having true values for all items in X and Y to the number of records having true values for all items in X. Support of the rule is the ratio of the number of records having true values for all items in X and Y to the number of all records in the database. The problem of association rule learning is to find all rules that satisfy the minimum support and minimum confidence constraints.

Association rule learning was applied in medicine, for example, to identify new and interesting patterns in surveillance data, in particular in the analysis of the *Pseudomonas aeruginosa* infection control data (Brossette *et al*., 1998). An algorithm for finding a more expressive variant of association rules, where data and patterns are represented in first-order logic, was successfully applied to the problem of predicting whether chemical compounds are carcinogenic or not (Toivonen and King, 1998).

Subgroup discovery (Wrobel, 1997, Gamberger and Lavrač, 2002, Lavrač *et al*., 2004) has the goal to uncover characteristic properties of population subgroups by building short rules which are highly significant (assuring that the distribution of classes of covered instances are statistically significantly different from the distribution in the training set) and have a large coverage (covering many target class instances). The approach, using a beam search rule learning algorithm aimed at inducing short rules with large coverage, was successfully applied to the problem of coronary heart disease risk group detection (Gamberger *et al*., 2003).

Genetic algorithms (Goldberg, 1989) are optimization procedures that maintain candidate solutions encoded as strings (or chromosomes). A fitness function is defined that can assess the quality of a solution represented by some chromosome. A genetic algorithm iteratively selects best chromosomes (i.e., those of highest fitness) for reproduction, and applies crossover and mutation operators to search in the problem space. Most often, genetic algorithms are used in combination with some classifier induction technique or some schema for classification rules in order to optimize their performance in terms of accuracy and complexity (e.g., (Larranaga *et al*., 1997) and (Dybowski *et al*., 1996)). They can also be used alone, e.g., for the estimation of Doppler signals (Gonzalez *et al*., 1999) or for multi-disorder diagnosis (Vinterbo and Ohno-Machado 1999). For more information please refer to Chapter 19 in this book.

Data analysis approaches reviewed so far in this chapter mostly use crisp logic: the attributes take a single value and when evaluated, decision rules return a single class value. *Fuzzy logic* (Zadeh, 1965) provides an enhancement compared to classical AI approaches (Steinmann, 1997): rather than assigning an attribute a single value, several values can be assigned, each with its own degree or grade. Classically, for example, "body temperature" of 37.2°C can be represented by a discrete value "high", while in fuzzy logic the same value can be represented by two values: "normal" with degree 0.3 and "high" with degree 0.7. Each value in a fuzzy set (like "normal" and "high") has a corresponding membership function that determines how the degree is computed from the actual continuous value of an attribute. Fuzzy systems may thus formalize a gradation and may allow handling of vague concepts—both being natural characteristics of medicine (Steinmann, 1997)—while still supporting comprehensibility and transparency by computationally relying on a fuzzy rules. In medical data analysis, the best developed approaches are those that use data to induce a straightforward tabular rule-based mapping from input to control variables and to find the corresponding membership functions. Example applications studies include design of patient monitoring and alarm system (Becker and Thull, 1997), support system for breast cancer diagnosis (Kovalerchuk *et al*., 1997), design of a rule-based visuomotor control (Prochazka, 1996). Fuzzy logic control applications in medicine are discussed in (Rau *et al*., 1995).

Support vector machines (SVM) are a classification technique originated from statistical learning theory (Cristianini, 2000, Vapnik, 1998). Depending on the chosen kernel, SVM selects a set of data examples (support vectors) that define the decision boundary between classes. SVM have been proven for excellent classification performance, while it is arguable whether support vectors can be effectively used in communication of medical knowledge to the domain experts.

Bayesian networks (Pearl, 1988) are probabilistic models that can be represented by a directed graph with vertices encoding the variables in the model and edges encoding their dependency. Given a Bayesian network, one can compute any joint or conditional probability of interest. In terms of intelligent data analysis, however, it is the learning of the Bayesian network from data that is of major importance. This includes learning of the structure of the network, identification and inclusion of hidden nodes, and learning of conditional probabilities that govern the networks (Szolovits, 1995, Lam, 1998). The data analysis then reasons about the structure of the network (examining the inter-variable dependencies) and the conditional probabilities (the strength and types of such dependencies). Examples of Bayesian network learning for medical data analysis include a genetic algorithm-based construction of a Bayesian network for predicting the survival in malignant skin melanoma (Larranaga *et al*., 1997), learning temporal probabilistic causal models from longitudinal data (Riva and Bellazzi, 1996), learning conditional probabilities in modeling of the clinical outcome after bone marrow transplantation (Quaglini *et al*., 1994), cerebral modeling (Labatut *et al*., 2003) and cardiac SPECT image interpretation (Sacha *et al*., 2002).

There are also different forms of unsupervised learning, where the input to the learner is a set of unclassified instances. Besides unsupervised learning using neural networks described in Section 58.3.2 and learning of association rules described in Section 58.4, other forms of unsupervised learning include conceptual clustering (Fisher, 1987, Michalski and Stepp, 1983) and qualitative modeling (Bratko, 1989).

The *data visualization techniques* may either complement or additionally support other data analysis techniques. They can be used in the preprocessing stage (e.g., initial data analysis and feature selection) and the postprocessing stage (e.g., visualization of results, tests of performance of classifiers, etc.). Visualization may support the analysis of the classifier and thus increase the comprehensibility of discovered relationships. For example, visualization of results of naive Bayesian classification may help to identify which are the important factors that speak for and against a diagnosis (Zelic *et al.*, 1997), and a 3D visualization of a decision tree may assist in tree exploration and increase its transparency (Kohavi *et al*., 1997).

58.5 Conclusions

There are many Data Mining methods from which one can chose for mining the emerging medical data bases and repositories. In this chapter, we have reviewed most popular ones, and gave some pointers where they have been applied. Despite the potential and promising approaches, the utility of Data Mining methods to analyze medical data sets is still sparse, especially when compared to classical statistical approaches. It is gaining ground, however, in the areas where data is accompanied with knowledge bases, and where data repositories storing heterogenous data from different sources took ground.

Acknowledgments

This work was supported by the Slovenian Ministry of Education, Science and Sport. Thanks to Elpida Keravnou, Riccardo Bellazzi, Peter Flach, Peter Hammond, Jan Komorowski, Ramon M. Lopez de Mantaras, Silvia Miksch, Enric Plaza and Claude Sammut for their comments on individual parts of this chapter.

References

Aamodt, A. and Plaza, E., Case-based reasoning: Foundational issues, methodological variations, and system approaches, *AI Communications*, 7(1): 39–59 (1994).

Agrawal, R., Manilla, H., Srikant, R., Toivonen, H. and Verkamo A.I., "Fast discovery of association rules." In: Advances in Knowledge Discovery and Data Mining (Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., eds.), AAAI Press, 1996, pp. 307–328 (1996).

Aha, D., Kibler, D., Albert, M., "Instance-based learning algorithms," *Machine Learning*, 6(1): 37–66 (1991).

Andrews, R., Diederich, J. and Tickle, A.B., "A survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge Based Systems*, 8(6): 373–389 (1995).

Andrews, P.J., Sleeman, D.H., Statham, P.F., *et al.* "Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression." *J Neurosurg.* 97(2): 326-336 (2002).

Astion, M.L. and Wilding, P., "The application of backpropagation neural networks to problems in pathology and laboratory medicine," *Arch Pathol Lab Med*, 116(10): 995–1001 (1992).

Averbuch, M., Karson, T., Ben-Ami, B., Maimon, O., and Rokach, L. (2004). Context-sensitive medical information retrieval, MEDINFO-2004, San Francisco, CA, September. IOS Press, pp. 282-262.

Barlow, T.W., "Self-organizing maps and molecular similarity," *Journal of Molecular Graphics*, 13(1): 53–55 (1995).

Baxt, W.G. "Application of artificial neural networks to clinical medicine," *Lancet*, 364(8983) 1135–1138 (1995).

Becker, K., Thull, B., Kasmacher-Leidinger, H., Stemmer, J., Rau, G., Kalff, G. and Zimmermann, H.J. "Design and validation of an intelligent patient monitoring and alarm system based on a fuzzy logic process model," *Artificial Intelligence in Medicine*, 11(1): 33–54 (1997).

Bradburn, C., Zeleznikow, J. and Adams, A., "Florence: synthesis of case-based and model-based reasoning in a nursing care planning system," *Computers in Nursing*, 11(1): 20–24 (1993).

Bratko, I., Kononenko, I. Learning diagnostic rules from incomplete and noisy data. In Phelps, B. (ed.) *AI Methods in Statistics*. Gower Technical Press, 1987.

Bratko, I., Mozetič, I. and Lavrač, N., *KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems*, The MIT Press, 1989.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

Brossette, S.E., Sprague, A.P., Hardin, J.M., Waites, K.B., Jones, W.T., Moser, S.A. "Association rules and Data Mining in hospital infection control and public health surveillance." *Journal of the Americal Medical Inform. Assoc.* 5(4): 373–81 (1998).

Breault, J.L., Goodall, C.R., Fos P.J., *Data Mining a Diabetic Data Warehouse*. Artificial Intelligence in Medicine 26(1-2): 37–54 2002.

Carpenter, G.A. and Tan, A.H., "Rule extraction, fuzzy artmap and medical databases." In: *Proc. World Cong. Neural Networks*, pp. 501–506 (1993).

Carrault, G., Cordier, M., Quiniou, R., Wang, F., *Temporal Abstraction and Inductive Logic Programming for Arrhythmia Recognition from Electrocardiograms*. Artificial Intelligence in Medicine 28(3): 231–236 (2003).

Caruana, R., Baluja, S., and Mitchell, T., "Using the Future to Sort Out the Present: Rankprop and Multitask Learning for Medical Risk Analysis," *Advances in Neural Information Processing Systems (NIPS*95)* 8: 959–965 (1995).

Cendrowka, J. "PRISM: An algorithm for inducing modular rules," *Int. J. Man-Machine Studies* 27: 349–370 (1987).

Cestnik B., "Estimating Probabilities: A Crucial Task in Machine Learning," In: *Proc. European Conf. on Artificial Intelligence*, pp. 147-149 (1990).

Cestnik B., Kononenko I., Bratko I., "ASSISTANT 86: A knowledge elicitation tool for sophisticated users." In: *Progress in Machine learning* (Bratko, I., Lavrač, N., eds.), Wilmslow: Sigma Press (1987).

Chung, F.L. and Lee, L. "A node prunning algorithm for backpropagation network," *Int. J. Neural Systems*, 3: 301–314 (1992).

Clark, P., Boswell, R., "Rule induction with CN2: Some recent improvements." In: *Proc. Fifth European Working Session on Learning*, Springer, pp. 151–163 (1991).

Clark, P., Niblett, T. The CN2 induction algorithm. *Machine Learning*, 3(4): 261–283 (1989).

Compton, P. and Jansen, R., "Knowledge in context: A strategy for expert system maintenance." In: *Proc. 2nd Australian Joint Artificial Intelligence Conference*, Springer LNAI 406, pp. 292–306 (1988).

Compton, P., Horn, R., Quinlan, R. and Lazarus, L., "Maintaining an expert system." In: *Applications of Expert Systems* (Quinlan, R., ed.), Addison Wesley, pp. 366–385 (1989).

Cover, T.M., Hart, P.E., "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, 13: 21–27 (1968).

Craven, M.W., and Shavlik, J.W. (1997) "Using neural networks for Data Mining," *Future generation computer systems*, 13(2–3): 211–229 (1997).

Cristianini, N., and Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge Univ. Press (2000).

Dasarathy, B.V. (ed.) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos, CA (1990).

Dehaspe, L, Toivonen, H. and King, R.D. "Finding frequent substructures in chemical compounds." In: *Proc. 4th International Conference on Knowledge Discovery and Data Mining, (KDD-98)* (Agrawal, R., Stolorz, P. and Piatetsky-Shapiro, G., eds.), AAAI Press, pp. 30–37 (1998)..

De Raedt, L. and Dehaspe, L., "Clausal discovery." *Machine Learning*, 26: 99–146 (1997).

Di Bona, S., Niemann, H., Pieri, G., Salvetti, O., *Brain Volumes Characterisation Using Hierarchical Neural Networks*. AI in Medicine 28 (2003).

Downs, J., Harrison, R.F., Kennedy, R.L., and Cross, S.C., "Application of the fuzzy artmap neural network model to medical pattern classification tasks," *Artificial Intelligence in Medicine*, 8(4): 403–428 (1996).

Dudani, S.A., "The distance-weighted *k*-nearest neighbor rule," *IEEE Transactions on Systems, Man and Cybernetics*, 6(4): 325–327 (1975).

Dybowski R., Weller P., Chang R., Gant V. "Prediction of outcome in the critically ill using an artificial neural network synthesised by a genetic algorithm," *Lancet*, 347: 1146-1150 (1996).

Džeroski, S., Lavrač, N., "Rule induction and instance-based learning applied in medical diagnosis," *Technology and Health Care*, 4(2): 203–221 (1996).

Edwards, G., Compton, P., Malor, R., Srinivasan, A. and Lazarus, L., "PEIRS: A pathologist maintained expert system for the interpretation of chamical pathology reports," *Pathology* 25: 27–34 (1993).

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, 39(11):27–41 (1996).

Fausett, L.V., *Fundamentals of neural networks: Architectures, algorithms and applications*, Prentice Hall, Upper Saddle River, NJ (1994).

Fisher, D.H. (1987) Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2: 139–172 (1987).

Fix, E., Hodges, J.L., "Discriminatory analysis. Nonparametric discrimination. Consistency properties." Technical Report 4, US Air Force School of Aviation Medicine. Randolph Field, TX (1957).

Frawley, W., Piatetsky-Shapiro, G., Matheus, C. "Knowledge discovery in databases: An overview." In: *Knowledge discovery in databases* (Piatetsky-Shapiro, G., Frawley, W., eds.), The AAAI Press, Menlo Park, CA (1991).

Gamberger D. and Lavrač N., Expert-guided subgroup discovery: Methodology and application. *Journal of Artficial Intelligence Research* 17: 501–527 (2002).

Gamberger, D., Lavrač, N., & Krstačić, G. Active subgroup mining: A case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine* 28(1): 27–57 (2003).

Gaspari, M., Roveda, G., Scandellari, C., Stecchi, S., *An Expert System for the Evaluation of EDSS in Multiple Sclerosis*. Artificial Intelligence in Medicine 25(2): 187–210, (2001).

Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*, pp. 68–74, Addison-Wesley (1989).

Gonzalez, J.S., Rodrigez, K., Garcia Nocetti, D.F., *Model-based Spectral Estimation of Doppler Signals Using Parallel Genetic Algorithms*. Artificial Intelligence in Medicine 19(1): 75–89 (1999).

Grzymała-Busse, J., "Applications of the rule induction systems LERS," In: (Polkowski and Skowron, 1998a), pp. 366–375 (1998).

Gyöngyösi, M., Ploner, M., Porenta, G., Sperker, W., Wexberg, P., Strehblow, C., Glogar, D., *Case-based Distance Measurements for the Selection of Controls in Case-matched studies: Application in Coronary Interventions*. Artificial Intelligence in Medicine 26(3): 237–53 (2002).

Ham, F.M. and Han, S. "Classification of cardiac arrhythmias using fuzzy artmap," *IEEE Transactions on Biomedical Engineering*, 43(4): 425–430 (1996).

Henson, D.B. , Spenceley, S.E., and Bull, D.R. "Artificial neural network analysis of noisy visual field data in glaucoma," *Artificial Intelligence in Medicine*, 10(2): 99–113 (1997).

Holte, R., Acker, L., Porter, B. "Concept learning and the problem of small disjuncts." In: *Proc. Tenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA, pp. 813–818 (1989).

Horn, K., Compton, P.J., Lazarus, L. and Quinlan, J.R. "An expert system for the interpretation of thyroid assays in a clnical laboratory," *Austr. Comput. Journal* 17(1): 7–11 (1985).

Kahn, C.E. Jr. and Anderson, G. M., "Case-based reasoning and imaging procedure selection," *Invest Radiol.*, 29(6): 643-647 (1994).

Kattan, M.W. and Beck, J.R., "Artificial neural networks for medical classification decisions," *Arch Pathol Lab Med*, 119: 672–677 (1995).

Kattan, M.W., Ishida, H., Scardino, P.T. and Beck, J.R., "Applying a neural network to prostate cancer survival data." In: *Intelligent data analysis in medicine and pharmacology* (Lavrač, N. Keravnou, E. and Zupan, B., eds.), Kluwer, pp. 295–306 (1997).

Kira, K., Rendell, L. "A practical approach to feature selection." In: *Proc. Intern. Conf. on Machine Learning* (Sleeman, D., Edwards, P., eds.), Aberdeen, Morgan Kaufmann, pp. 249–256 (1992).

Kira, K., Rendell, L. "The feature selection problem: traditional methods and new algorithm." In: *Proc. AAAI'92*, San Jose, pp. 129–134 (1992).

Koehle, M., Merkl, D., Kastner, J. "Clinical Gait Analysis by Neural Networks - Issues and Experiences." In: *Proc. IEEE Symposium on Computer-Based Medical Systems (CBMS'97)* (Kokol, P., Štiglič, B., eds.), Maribor, IEEE Press, pp. 138–143 (1997).

Kohavi, R., Sommerfield, D. and Dougherty, J., "Data Mining using MLC++, a machine learning library in C++," *International Journal of Artificial Intelligence Tools*, 6(4): 537–566 (1997).

Kohonen, T., *Self-organization and associative memory*, Springer-Verlag, New York (1988).

Komorowski, J. and Øhrn, A., "Modelling prognostic power of cardiac tests using rough sets," *Artificial Intelligence in Medicine*, 15(2): 167–91 (1998).

Kononenko, I., "Semi-naive Bayesian classifier." In: *Proc. European Working Session on Learning* (Kodratoff, Y., ed.), Porto, Springer, pp. 206–219 (1991).

Kononenko, I., "Inductive and Bayesian learning in medical diagnosis," *Applied Artificial Intelligence*, 7: 317–337 (1993).

Kononenko, I. (1994) "Estimating attributes: Analysis and extensions of Relief." In: *Proc. European Conf. on Machine Learning* (De Raedt, L., Bergadano, F., eds.), Catania, Springer, pp. 171–182 (1994).

Kononenko, I., Šimec, E. (1995) "Induction of decision trees using RELIEFF." In: *Proc. of ISSEK Workshop on Mathematical and Statistical Methods in Artificial Intelligence* (Della Riccia, G., Kruse, R., Viertl, R., eds.), Udine, September 1994, Springer, pp. 199–220 (1995).

Kovalerchuk, B., Triantaphyllou, E., Ruiz, J.F. and Clayton, J., "Fuzzy logic in computer-aided breast cancer diagnosis: analysis of lonulation," *Artificial Intelligence in Medicine*, 11(1): 75–87 (1997).

Labatut, V., Pastor, J., Ruff, S., Demonet, J.F., Celsis, P., *Cerebral Modeling and Dynamic Bayesian Networks*. *Artificial Intelligence in Medicine* 30(1): 119–39 (2003).

Lam, W. (1998), Bayesian network refinement via machine learning approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3): 240–251 (1998).

Larranaga, P., Sierra, B., Gallego, M.J., Michelena, M.J., Picaza, J.M., Learning Bayesian networks by genetic algorithms: a case study in the prediction of survival in malignant skin melanoma. In *Proc. Artificial Intelligence in Medicine Europe* (E. Keravnou, C. Garbay, R. Baud, J. Wyatt, eds.), pp. 261–272 (1997).

Lavrač, N., Džeroski, S., Pirnat, V., Križman, V. "The utility of background knowledge in learning medical diagnostic rules," *Applied Artificial Intelligence*, 7: 273–293 (1993).

Lavrač, N., Džeroski, S., *Inductive Logic Programming: Techniques and Applications.* Ellis Horwood, Chichester (1994).

Lavrač, N., Kavšek, B., Flach, P., and Todorovski, L. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5: 153–188, 2004.

Liestøl, K., Andersen, P.K. and Andersen, U. "Survival analysis and neural nets," *Statist. Med.*, 13(2): 1189–1200 (1994).

Lin, T.Y and Cercone, N., eds., "Rough Sets and Data Mining", Kluwer (1997).

Lubsen, J., Pool, J., van der Does, E. A practical device for the application of a diagnostic or prognostic function. *Methods Inf. Med.* 17(2): 127–129 (1978).

Macura, R.T. and Macura, K., eds., "Case-based reasoning: opportunities and applications in health care," *Artificial Intelligence in Medicine*, 9(1): 1–4 (1997).

Macura, R.T. and Macura, K., eds., *Artificial Intelligence in Medicine: Special Issue on Case-Based Reasoning*, 9(1) (1997).

Malmberg, L.P., Kallio, K., Haltsonen, S., Katila, T. and Sovijarvi, A.R., "Classification of lung sounds in patients with asthma, emphysema, fibrosing alveolitis and healthy lungs by using self-organizing maps," *Clinical Physiology*, 16(2): 115–129 (1996).

Mariuzzi, G., Mombello, A., Mariuzzi, L., Hamilton, P.W., Weber, J.E., Thompson D. and Bartels, P.H., "Quantitative study of ductal breast cancer–patient targeted prognosis: an exploration of case base reasoning," *Pathology, Research & Practice*, 193(8): 535–542 (1997).

Markey, M.K., Lo, J.Y., Tourassi, G.D. AND Floyd Jr., C.E., "Self-organizing map for cluster analysis of a breast cancer database," *Artificial Intelligence in Medicine*, 27(2): 113-127 (2002).

McSherry, D., "Hypothesist: A development environment for intelligent diagnostic systems." In: *Proc. Sixth Conference on Artificial Intelligence in Medicine* (AIME'97), Springer, pp. 223–234 (1997).

Michalski, R.S. and Stepp, R.E., "Learning from observation: Conceptual clustering." In: *Machine Learning: An AI Approach* (Michalski, R.S., Carbonell, J. and Mitchell, T.M., eds.), volume I, Palo Alto, CA. Tioga., pp. 331–363 (1983).

Michalski, R.S. (1986) "Understanding the nature of learning: Issues and research directions." In: *Machine Learning: An AI Approach* (Michalski, R.S., Carbonnel, J. and Mitchell, T.M., eds.) Morgan Kaufmann, pp. 3–25 (1986).

Michalski, R.S., Mozetič, I., Hong, J. and Lavrač, N., "The multi-purpose incremental learning system AQ15 and its testing application on three medical domains." In *Proc. Fifth National Conference on Artificial Intelligence*, Morgan Kaufmann, 1986, pp. 1041–1045.

Mizoguchi, F., Ohwada, H., Daidoji, M., Shirato, S., "Using Inductive Logic Programming to learn classification rules that identify glaucomatous eyes." In: *Intelligent Data Analysis in Medicine and Pharmacology* (Lavrač, N., Keravnou, E., Zupan, B., eds.), Kluwer, pp. 227–242 (1997).

Modai, I., Israel, A., Mendel, S., Hines, E.L. and Weizman, R., "Neural network based on adaptive resonance theory as compared to experts in suggesting treatment for schizophrenic and unipolar depressed in-patients," *Journal of Medical Systems*, 20(6): 403–412 (1996).

Muggleton, S., "Inverse entailment and Progol," *New Generation Computing, Special Issue on Inductive Logic Programming*, 13(3–4): 245–286 (1995).

Niblett, T. and Bratko, I., "Learning decision rules in noisy domains." In: *Research and Development in Expert Systems III* (Bramer, M., ed.), Cambridge University Press, pp. 24–25 (1986).

Pawlak, Z., Information systems – theoretical foundations. *Information Systems*, 6: 205–218 (1981).

Pawlak, Z., *Rough Sets: Theoretical Aspects of Reasoning about Data*, volume 9 of *Series D: System Theory, Knowledge Engineering and Problem Solving*. Kluwer (1991).

Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA (1988).

Pilih, I.A., Mladenič, D., Lavrač, N., Prevec, T.S., "Data analysis of patients with severe head injury." In: *Intelligent Data Analysis in Medicine and Pharmacology* (Lavrač, N., Keravnou, E., Zupan, B., eds.), Kluwer, pp. 131–148 (1997).

Polkowski, L. and Skowron, A., eds., *Rough Sets in Knowledge Discovery 1: Methodology and Applications*, volume 18 of *Studies in Fuzziness and Soft Computing*. Physica-Verlag (1998).

Polkowski, L. and Skowron, A., eds. (1998) *Rough Sets in Knowledge Discovery 2: Applications, Case Studies and Software Systems*, volume 18 of *Studies in Fuzziness and Soft Computing*. Physica-Verlag (1998).

Prochazka, A., "The fuzzy logic of visuomotor control," *Canadian Journal of Physiology & Pharmacology*, 74(4): 456–462 (1996).

Quaglini, S., Bellazzi, R., Locatelli, F., Stefanelli, M., Salvaneschi, C., "An Influence Diagram for Assessing GVHD Prophylaxis after Bone Marrow Transplantation in Children." *Medical Decision Making*, 14:223-235 (1994).

Quinlan, J.R., "Induction of decision trees." *Machine Learning*, 1(1): 81–106 (1986).

Quinlan, J.R., "Learning logical definitions from relations," *Machine Learning*, 5(3): 239–266 (1990).

Quinlan, J.R., *C4.5: Programs for Machine Learning*, San Mateo, CA, Morgan Kaufmann (1993).

Rau, G., Becker, K., Kaufmann, R. and Zimmermann, H.J., Fuzzy logic and control: principal approach and potential applications in medicine, *Artificial Organs*, 19(1): 105–112 (1995).

Richeldi, M., Rossotto, M., "Class-driven statistical discretization of continuous attributes." In: *Machine Learning: Proc. ECML-95* (Lavrač, N., Wrobel, S., eds.), Springer, pp. 335-342 (1995).

Riva, A. and Bellazzi, R., "Learning Temporal Probabilistic Causal Models from Longitudinal Data." *Artificial Intelligence in Medicine*, 8(3): 217–234 (1996).

Rokach, L., Averbuch, M., and Maimon, O., Information retrieval system for medical narrative reports (pp. 217228). Lecture notes in artificial intelligence, 3055. Springer-Verlag (2004).

Rosetta: A rough set toolkit for the analysis of data.

Rumelhart, D.E. and McClelland, J.L., eds., *Parallel Distributed Processing, Vol. 1: Foundations*. MIT Press, Cambridge, MA (1986).

Sacha, J.P., Goodenday, L.S. and Cios, K.J., "Bayesian Learning for Cardiac SPECT Image Interpretation". AI in Medicine 26(1): 109–143 (2002).

Sammut, C., "Introduction to Ripple Down Rules." `http://www.cse.unsw.edu.au/~claude/teaching/AI/notes/prolog/Extenions/rdr.lens.html` (1998).

Setiono, R., "Extracting rules from pruned networks for breast cancer diagnosis," *Artificial Intelligence in Medicine*, 8(1): 37–51 (1996).

Setiono, R. "Extracting rules from neural networks by pruning and hidden-unit splitting," *Neural Computation*, 9(1): 205–225 (1997).

Setiono, R. "Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis" *Artificial Intelligence in Medicine*, 18(3): 205–219, (1999).

Shannon, C.E., "A mathematical theory of communication." *Bell. Syst. Techn. J.*, 27: 379–423 (1948).

Shawlik, J.W., Mooney, R.J. and Towell, G.G., "Symbolic and neural learning algorithms: An experimental comparison," *Machine Learning*, 6(2): 111–143 (1991).

Skowron, A., "Synthesis of adaptive decision systems from experimantal data (invited talk)." In: *Proc. of the Fifth Scandinavian Conference on Artificial Intelligence SCAI-95* (A. Aamodt and J. Komorowski, eds.), IOS Press Ohmsa, Amsterdam, pp. 220–238 (1995).

Stel, V.S., Pluijm, S.M., Deeg, D.J., Smit, J.H., Bouter, L.M., Lips, P. "A classification tree for predicting recurrent falling in community-dwelling older persons." *J Am Geriatr Soc.* 51(10): 1356–1364 (2003).

Srinivasan, A., King, R.D., Muggleton, S.H. and Sternberg, M.J.E., "Carcinogenesis predictions using inductive logic programming." In *Intelligent Data Analysis in Medicine and Pharmacology* (Lavrač, N. Keravnou, E. and Zupan, B., eds.), Kluwer, pp. 243–260 (1997)..

Steinmann, F., "Fuzzy set theory in medicine," *Artificial Intelligence in Medicine*, 11(1) 1–7 (1997).

Szolovits, P., "Uncertainty and Decision in Medical Informatics," *Methods of Information in Medicine*, 34: 111–121 (1995).

Tsumoto, S., "Modelling medical diagnostic rules based on rough sets", In: *Proc. First International Conference on Rough Sets and Soft Computing – RSCTC'98* (Polkowski, L. and Skowron, A., eds.), volume 1424 of *Lecture Notes in Artificial Intelligence, Springer Verlag*. Springer, pp. 475–482 (1998).

Vapnik, V.N. *Statistical Learning Theory*. Wiley (1998).

Vinterbo, S. and Ohno-Machado, L., *A Genetic algorithm approach to multi-disorder diagnosis*. *AI in Medicine*, 18(2): 117–32 (1999).

Weiss, S.M., Kulikowski, C.A., *Computer Systems that Learn*. Morgan Kaufmann, San Mateo, CA, 1991.

Wettschereck, D., "A study of distance-based machine learning algorithms," PhD Thesis, Department of Computer Science, Oregon State University, Corvallis, OR (1994).

Witten, I. H. and E. Frank, *Data Mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, CA (1999).

Wolpert, D., "Constructing a generalizer superior to NETtalk via mathematical theory of generalization," *Neural Networks*, 3: 445–452 (1989).

Wrobel, S., "An algorithm for multi-relational discovery of subgroups." In: *Proc. First European Symposium on Principles of Data Mining and Knowledge Discovery*, Springer, pp. 78–87 (1997).

Wu, C.H., Artificial neural networks for molecular sequence analysis, *Computers & Chemistry*, 21(4): 237–56 (1997).

Zadeh, L.A., Fuzzy sets. *Information and Control*, vol. 8, pp. 338–353 (1965).

Zelič, I., Kononenko, I., Lavrač, N., Vuga, V., "Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries," *Journal of Medical Systems*, 21(6): 429–444 (1997).

Zhu, Y. and Yan, H., "Computerized tumor boundary detection using a hopfield neural network," *IEEE Transactions on Medical Imaging*, 16(1): 55–67 (1997).

Zupan, B. and Džeroski, S., "Acquiring and validating background knowledge for machine learning using function decomposition," *Artificial Intelligence in Medicine*, 14(1–2): 101–118 (1998).

Zupan, B., Halter, J.A. and Bohanec, M. (1997) "Concept discovery by decision table decomposition and its application in neurophysiology." In *Intelligent Data Analysis in Medicine*

and Pharmacology (Lavrač, N., Keravnou, E. and Zupan, B., eds.), Kluwer, pp. 261–277 (1997).

Zupan, B., Bohanec, M., Demšar, J. and Bratko, I., "Feature transformation by function decomposition," *IEEE Intelligent Systems*, 13(2): 38–43 (1998).

Zupan, B., Demsar, J., Smrke, D., Bozikov, K., Stankovski, V., Bratko, I., Beck, J.R., "Predicting patient's long-term clinical status after hip arthroplasty using hierarchical decision modelling and Data Mining." *Methods Inf Med*. 40(1): 25–31 (2001).