Rule induction and instance-based learning applied in medical diagnosis

Article *in* Technology and health care: official journal of the European Society for Engineering and Medicine · September 1996

DOI: 10.3233/THC-1996-4208 · Source: PubMed

2 authors:

Sašo Džeroski
Jožef Stefan Institute
549 PUBLICATIONS **14,635** CITATIONS
SEE PROFILE

Nada Lavrac
Jožef Stefan Institute
409 PUBLICATIONS **10,230** CITATIONS
SEE PROFILE

Technology and Health Care 4 (1996) 203–221
IOS Press

Rule induction and instance-based learning applied in medical diagnosis

Sašo Džeroski and Nada Lavrač
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
Tel.: +386 61 177 3 217; Fax: +386 61 125 1 038
E-mail: Saso.Dzeroski@ijs.si, Nada.Lavrac@ijs.si

Paper honouring the memory of Tim de Dombal

Abstract. Machine learning methods have been applied in a variety of medical domains in order to improve medical decision making. Improved medical diagnosis and prognosis can be achieved through automatic analysis of patient data stored in medical records, i.e., by learning from past experience. Given patient records with corresponding diagnoses, machine learning methods are able to classify new cases either through constructing explicit rules that generalize the training cases (e.g., rule induction) or by storing (some of) the training cases for reference (instance-based learning). This paper presents the methodologies of rule induction and instance-based learning and their application to medical diagnosis, in particular, the problem of early diagnosis of rheumatic diseases. It also discusses the possibility to use existing expert knowledge to support the learning process and the utility of such knowledge.

Keywords: Machine learning, instance-based learning, rule induction, medical diagnosis, rheumatic diseases

1. Introduction

Current trends in medical decision making show awareness of the need of introducing formal techniques, as well as intelligent data analysis techniques that enable the extraction of knowledge, regularities, trends and representative cases from patient data stored in medical records. Formal techniques include decision theory [13] and symbolic reasoning technology [22], as well as methods at their intersection, such as probabilistic belief networks [30]. Intelligent data analysis techniques include machine learning, clustering, data visualization, and interpretation of time-ordered data (derivation and revision of temporal trends and other forms of temporal data abstraction).

This paper is concerned with methods for intelligent data analysis in medicine, in particular machine learning methods [23, 25]. Machine learning methods can be classified into three major groups [25]: inductive learning of symbolic rules (such as induction of rules [6, 24], decision trees [32] and induction of logic programs [18]), statistical or pattern-recognition methods (such as k-nearest neighbors or instance-based learning [1, 8], discriminate analysis and Bayesian classifiers), and artificial neural networks [33] (such as networks with backpropagation learning, Kohonen's self-organizing network and Hopfield's associative memory).

The importance of machine learning methods is due to the fact that the gap between data generation/storage and data comprehension is widening in all fields of human activity, also in medicine. Overcoming this gap is crucial for improving performance. Thus, medical decision making needs

to be supported by arguments based on basic medical and pharmacological knowledge as well as knowledge extracted from data by techniques of machine learning in the form of regularities, trends and typical cases.

Machine learning methods have been applied in a variety of medical domains in order to improve medical decision making [17]. These include diagnostic and prognostic problems in oncology [2], liver pathology [21], neuropsychology [27], and gynaecology [29]. Improved medical diagnosis and prognosis may be achieved through automatic analysis of patient data stored in medical records, i.e., by learning from past experiences.

Given patient records with corresponding diagnoses, machine learning methods are able to diagnose new cases. More specifically, given is a set of examples with known classifications. An example is described by the values of a fixed collection of features (attributes): A_i, $i \in \{1, \ldots, n\}$. Each attribute can either have a finite set of values (discrete) or take real numbers as values (continuous). A particular example e_j is thus a vector of attribute values: $e_j = (v_{1j}, \ldots, v_{nj})$. Each example is assigned one of N possible values of the class C (classifications): c_i, $i \in \{1, \ldots, N\}$. The class of example e_j will be denoted by c_j. For instance, in the domain of early diagnosis of rheumatic diseases, described in detail in Section 4, the patient records comprise 16 anamnestic attributes. Some of these are continuous (e.g., age, duration of morning stiffness) and some are discrete (e.g., joint pain, which can be arthrotic, arthritic, or not present at all). There are eight possible diagnoses: degenerative spine diseases, degenerative joint diseases, inflammatory spine diseases, other inflammatory diseases, extra-articular rheumatism, crystal-induced synovitis, non-specific rheumatic manifestations, and non-rheumatic diseases.

To classify (diagnose) new cases, machine learning methods can take different approaches. One is to construct explicit symbolic rules that generalize the training cases (rule induction). The induced rules can then be used to classify new cases. Another approach is to store (some of) the training cases for reference (instance-based learning). New cases can then be classified by comparing them to the reference cases.

This paper presents the methodologies of rule induction [6, 23] and instance-based learning [1, 8, 37] and their application to medical diagnosis. It describes in detail the rule induction algorithm CN2 [5, 10] (Section 2) and the k-nearest neighbor-based learning algorithm of Wettschereck [37] (Section 3). It then describes the problem of early diagnosis of rheumatic diseases [15, 20, 31] and the application of the two machine learning methodologies to this problem (Section 4). The possibility to use existing expert knowledge to support the learning process and the utility of such knowledge is discussed in Section 5. Section 6 concludes with a discussion of the advantages and disadvantages of each on the two machine learning methodologies.

2. Rule induction with CN2

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

IF condition THEN conclusion.

The condition contains one or more attribute tests of the form $A_i = v_i$ for discrete attributes and $A_i < v_i$ or $A_i > v_i$ for continuous attributes. The conclusion part has the form $C = c_i$, assigning a particular value c_i to the class C. We say that an example is covered by a rule if the attribute values of the example obey the conditions in the IF part of the rule.

Table 1
An example if–then rule induced by CN2 in the domain of early diagnosis of rheumatic diseases

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

An example rule induced in the domain of early diagnosis of rheumatic diseases, described in detail in Section 4, is given in Table 1. It assigns the diagnosis of crystal-induced synovitis to male patients older than 46 that have more than three painful joints and psoriasis as a skin manifestation.

In our experiments, we used the rule induction system CN2 [5, 6, 10]. CN2 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 positive, all other examples are negative. The covering approach works as follows: CN2 constructs a rule that correctly classifies some examples, removes the positive examples covered by the rule from the training set and repeats the process until no more examples remain. To construct a single rule that classifies examples into class c_i, CN2 starts with a rule with an empty antecedent (IF part) and the selected class c_i as a consequent (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 the class c_i satisfy the antecedent. To allow for 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 already fixed 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 $A_i = v_i$, where v_i is a possible value for A_i, are considered for inclusion in the condition part. For continuous attributes, all conditions of the form $A_i < (v_{ik} + v_{i(k+1)})/2$ and $A_i > (v_{ik} + v_{i(k+1)})/2$ are considered, where v_{ik} 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 appear in the current training set, the conditions $A < 1.5$, $A > 1.5$, $A < 3.0$, and $A > 3.0$ 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 are mainly designed to estimate the performance of the rule on unseen examples in terms of classification accuracy. This is in accord with the task of achieving high classification accuracy on unseen cases.

Suppose a rule covers p positive and n negative examples. Its accuracy can be estimated by the relative frequency of positive examples covered, computed as $p/(p+n)$. This heuristic was used in early rule induction algorithms. It 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 only a few examples. In the extreme case, a maximally specific rule will cover (be supported by) one example and hence have an unbeatable score using the metrics of apparent accuracy (scores 100% accuracy). Apparent accuracy on the training data, however, does not adequately reflect true predictive accuracy, i.e., accuracy on new testing data. It has been shown [14] that rules supported by few examples have very high error rates on new testing data.

The problem lies in the estimation of the probabilities involved, i.e., the probability that a new example 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.

A more recent version of CN2 [5] uses the Laplace estimate to estimate the accuracy of rules. This estimate is more reliable than relative frequency. If a rule covers p positive and n negative examples, its accuracy is estimated as $(p+1)/(p+n+N)$, where N is the number of possible classes.

Unfortunately, the Laplace estimate relies on the assumption that all classes are equally probable a priori, an assumption which is rarely true in practice. We have therefore extended CN2 [10] to enable the use of an even more sophisticated probability estimate, i.e., the m-estimate [3]. The m-estimate takes into account the prior probabilities of each class, and combines them with the evidence provided by the examples covered by the particular rule. The parameter m controls the role of the prior probabilities and the evidence provided by the examples: higher m gives more weight to the prior probabilities and less to the examples. Higher values of m are thus appropriate for examples that contain more noise. If a rule that predicts class c_i covers p positive and n negative examples, its accuracy is estimated to be $(p+mp_i)/(p+n+m)$ [3], where p_i is the prior probability of class c_i. In CN2, p_i is estimated from the complete training set by relative frequency.

CN2 can also use a significance measure to enforce the induction of reliable rules. A rule is deemed reliable (significant) if the class distribution of the examples it covers is significantly different from the prior class distribution as given by the entire training set. This is measured by the likelihood ratio statistic [5]. Suppose the rule covers r_i examples of class c_i, $i \in \{1,\ldots,N\}$. Let $q_i = r_i/(r_1+\cdots+r_N)$ and let p_i be the prior probability of class c_i. The value of the likelihood ratio statistic is then

$$2(r_1+\cdots+r_N)\sum_{i=1}^{N} q_i \log_2(q_i/p_i).$$

This statistic is distributed as χ^2 with $N-1$ degrees of freedom. If its value is above a specified significance threshold, the rule is deemed significant.

CN2 can induce a set of if–then rules which is either ordered or unordered. In the first case, the rules are considered precisely in the order specified: given an example to classify, the class predicted by the first rule that covers the example is returned. In the second case, all rules are checked and all the rules that cover the example are taken into account. Conflicting decisions are resolved by taking into account the number of examples of each class (from the training set) covered by each rule. Suppose we have a two-class problem and two rules with coverage $(10,2)$ and $(4,40)$ apply, i.e., the first rule covers 10 examples of class c_1 and 2 examples of class c_2, while the second covers 4 examples of class c_1 and 40 examples of class c_2. The 'summed' coverage would be $(14,42)$ and the example is assigned class c_2. The recent version of CN2 [10] can give probabilistic classifications: in the example above, we divide the coverage $(14,42)$ with the total number of examples covered (56) and obtain as an answer the probability distribution $(0.25,0.75)$. This means that the probability of the example belonging to class c_1 is $1/4$, while for c_2 that probability is $3/4$.

Another feature of the latest version of CN2 [10] is the possibility to measure the *information score* [16] of induced rules. The information score is a performance measure which is not biased by the prior class distribution. It accounts for the possibility to achieve high accuracy easily in domains

with a very likely majority class: classifying into the majority class by returning the prior probability distribution all the time gives a zero information score.

Let the correct class of example e_k be c_k, its prior probability $P(c_k)$ and the probability returned by the classifier $P'(c_k)$. The information score of this answer is

$$I(e_k) = \begin{cases} -\log P(c_k) + \log P'(c_k), & P'(c_k) \geqslant P(c_k), \\ \log(1 - P(c_k)) - \log(1 - P'(c_k)), & P'(c_k) < P(c_k). \end{cases}$$

As $I(e_k)$ indicates the amount of information about the correct classification of e_k gained by the classifier's answer, it is positive if $P'(c_k) > P(c_k)$, negative if the answer is misleading ($P'(c_k) < P(c_k)$) and zero if $P'(c_k) = P(c_k)$.

The *relative information score* I_r of the answers of a classifier on a testing set consisting of examples $e_1, e_2, \ldots, e_t$ belonging to one of the classes $c_1, c_2, \ldots, c_N$ can be calculated as the ratio of the *average information score of the answers* and the *entropy of the prior distribution of classes*.

$$I_r = \frac{\frac{1}{t} \times \sum_{k=1}^{t} I(e_k)}{-\sum_{i=1}^{N} P(c_i) \times \log P(c_i)}.$$

CN2 handles examples that have missing values for some attributes in a relatively straightforward fashion. If an example has an unknown value of attribute A, it is not covered by rules that contain conditions that involve attribute A. Note that this example may be covered by rules that do not refer to attribute A in their condition part.

In our experiments, CN2 was used to induce sets of unordered rules. The rules were required to be highly significant (at the 99% level) and thus reliable. Except for the significance threshold and the search heuristic settings described below, the parameter settings of CN2 were the default ones (see [5]).

Both the Laplace estimate and the m-estimate were used: we give results for both of them. To select the appropriate value for the parameter m, we used the following methodology. Fifteen different values of the parameter m were tried (0, 0.01, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024), as suggested by earlier experiments [3, 10]. For a given learning problem, we thus induced 15 sets of rules and chose the best according to the relative information score on the training set. This procedure allows us to choose the right level of fitting: overfitting is prevented by applying the significance threshold. Given this methodology, experiments on the entire dataset of patient records were performed, as well as experiments on subsets of this dataset (see Sections 4 and 5). The best values of m ranged between 16 and 128 depending on the presence of background knowledge and the choice of subsets of cases for training and testing.

3. Instance-based learning

Instance-based learning (IBL) algorithms [1] use specific instances to perform classification tasks, rather than using generalizations 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 effort towards inductive generalization 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 [7, 12]. The nearest-neighbor (NN) algorithm is one of the best known classification algorithms and an enormous body of research exists on the subject (see, for example, [8]). 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 that example and each of the 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 [11]. An optimal value of k may be determined automatically from the training set by using leave-one-out cross-validation [35]. In our experiments, the best k from the range $(1, 75)$ was chosen in this manner.

Finally, the contribution of each attribute to the distance may be weighted, in order to avoid problems caused by irrelevant features [36]. The feature weights are determined on the training set by using one of a number of alternative feature-weighting methods. In our experiments, we used the k-NN algorithm as implemented by Wettschereck [37], which includes the improvements described above. A more detailed description of how distance computation, classification, and feature weighting is performed, is given below.

Given two examples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, their distance is calculated as

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

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

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

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

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

where x is one of the k nearest neighbors of z and x_{c_j} is 1 if x belongs to class c_j. The class c_j with the 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 an 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 weighting 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. The mutual information [34] $I(C, A)$ between the class C and attribute A is thus a natural quantity with which the feature A is weighted in the k-NN implementation of Wettschereck [37] that we employed in our experiments.

The mutual information [34] between two variables is defined as the reduction in uncertainty concerning the value of one variable that is obtained when the value of the other variable is known. If an attribute provides no information about the class, the mutual information will be zero. The mutual information between the random variables X and Y is defined as $I(X,Y) = H(X) - H(X|Y)$, where $H(X)$ is the entropy of the random variable X with probability mass function $P(x)$, defined as $H(X) = -\sum_x \log_2 P(x)$. For discrete X and Y, it can be also calculated as

$$I(X,Y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}.$$

For continuous variables, probability densities have to be used instead of probability masses and integrals instead of sums. The probabilities involved are in our case estimated from the training examples.

Unknown values in the examples are handled through a modification of the distance function between examples. Only features that have known values are used in calculating the distance, and the number of features for which both examples have known values is taken into account. The modified distance function is thus

$$\mathrm{d}(x,y) = \sqrt{\sum_{i=1}^{n} w_i \times \mathrm{diff}(x_i, y_i)^2} / \sqrt{\text{number of features } i \text{ for which both } x_i \text{ and } y_i \text{ are known}},$$

where $\mathrm{diff}(x_i, y_i) = 0$ if either x_i or y_i are unknown and $\mathrm{diff}(x_i, y_i) = \mathrm{difference}(x_i, y_i)$ otherwise.

4. Early diagnosis of rheumatic diseases

Correct diagnosis in the early stage of a rheumatic disease is a difficult problem. Having passed all the investigations, many patients cannot be reliably diagnosed after their first visit to the specialist. The reason is that anamnestic, clinical, laboratory and radiological data of patients with different rheumatic diseases are frequently similar. In addition, the diagnosis can also be incorrect due to the subjective interpretation of data [31].

4.1. Patient data

Data about 462 patients were collected at the University Medical Center in Ljubljana, Slovenia [31]. There are over 200 different rheumatic diseases which can be grouped into three, six, eight or twelve diagnostic classes. Eight diagnostic classes were considered, as in the experiments of Karalič and Pirnat [15]. Table 2 gives the names of the diagnostic classes and the numbers of patients belonging to each class.

To facilitate the comparison with earlier experiments in rule induction in this domain [20], the experiments were performed on anamnestic data, without taking into account data about patients'

Table 2
The eight diagnostic classes and the corresponding numbers of patients

Class	Name	Number of patients
$A1$	Degenerative spine diseases	158
$A2$	Degenerative joint diseases	128
$B1$	Inflammatory spine diseases	16
$B234$	Other inflammatory diseases	29
C	Extra-articular rheumatism	21
D	Crystal-induced synovitis	24
E	Non-specific rheumatic manifestations	32
F	Non-rheumatic diseases	54

clinical manifestations, laboratory and radiological findings. The sixteen anamnestic attributes are as follows: sex, age, family anamnesis, duration of present symptoms (in weeks), duration of rheumatic diseases (in weeks), joint pain (arthrotic, arthritic), number of painful joints, number of swollen joints, spinal pain (spondylotic, spondylitic), other pain (headache, pain in muscles, thorax, abdomen, heels), duration of morning stiffness (in hours), skin manifestations, mucosal manifestations, eye manifestations, other manifestations and therapy.

Out of 462 patient records, eight were incomplete; twelve attribute values were missing (for attributes sex and age). This was not problematic since both CN2 and k-NN can handle missing data. The data are very noisy, i.e., unreliable, for the following reasons:

- Anamnestic data are by nature very noisy since they are, in fact, patients' own description of the disease, only interpreted by a specialist for rheumatic diseases. Interpretation of this data is subjective and therefore extremely unreliable.
- The grouping of about 200 different diagnoses into only eight diagnostic classes is problematic. For degenerative diseases (classes $A1$ and $A2$) many examples are available. Nearly 74% of all the data set consists of patient records for these two diagnostic classes, together with the class of non-rheumatic diseases (see Table 2). Furthermore, some diagnostic classes are relatively non-homogeneous, having few common characteristics.
- Some of the patients had more than one diagnosis, but only one diagnosis was included in the example set.
- Data were collected by different medical doctors without achieving their collective consensus.

4.2. *Results of rule induction on the entire dataset*

In the first group of experiments, the data about all 462 patients were used. The Laplace and the m-estimate were used within CN2. A 99% significance threshold was applied within CN2. With the Laplace estimate, CN2 induced a set of 30 rules (102 conditions), with classification accuracy of 51.7% and relative information score of 22%. These results are taken from [20]. With the m-estimate, $m = 64$ proved to be the best value, yielding a set of 21 rules (102 conditions) with accuracy of 63.6% and a relative information score of 45%. There are fewer rules when the m-estimate is used (although rules are longer on the average) and much higher accuracy and relative information score are achieved. This indicates that the use of the m-estimate allows for a better fit to the training set in the presence of a high significance threshold.

A selection of the 21 rules (one for each diagnostic class) is given in Table 3. Take, for example, the second rule in Table 3: it assigns the diagnosis of degenerative joint diseases to a female patient

older than 47 with arthrotic joint pain, no more than 18 painful joints and no spinal pain. The rule is supported by 47 examples with that diagnosis, but also covers eight cases that were diagnosed otherwise (three with degenerative spine diseases, two with non-specific rheumatic manifestations, and three with non-rheumatic diseases): this is indicated by the numbers in the square brackets given after the diagnosis.

Table 3
A selection of rules for early diagnosis of rheumatic diseases induced with CN2 using the m-estimate

```
IF    Age < 55
  AND 3 < Duration_of_present_symptoms < 113
  AND Duration_of_rheumatic_diseases < 13
  AND Number_of_swollen_joints < 3
  AND Spinal_pain = spondylotic
  AND Duration_of_morning_stiffness < 1.25
  AND Skin_manifestations = no
THEN  Diagnosis = Degenerative_spine_diseases [58 4 0 0 2 0 1 3]

IF    Sex = female
  AND Age > 47
  AND Joint_pain = arthrotic
  AND Number_of_painful_joints < 19
  AND Spinal_pain = no
THEN  Diagnosis = Degenerative_joint_diseases [3 47 0 0 0 0 2 3]

IF    Sex = male
  AND Number_of_painful_joints < 3
  AND Spinal_pain = spondylitic
THEN  Diagnosis = Inflammatory_spine_diseases [9 0 12 1 0 0 0 1]

IF    Age < 67
  AND Number_of_painful_joints > 1
  AND Number_of_swollen_joints > 0
  AND Other_pain = no
  AND Duration_of_morning_stiffness > 0.35
  AND Skin_manifestations = no
  AND Eye_manifestations = no
THEN  Diagnosis = Other_inflammatory_diseases [2 2 0 11 0 0 0 0]

IF    20 < Age < 39
  AND 0 < Duration_of_present_symptoms < 21
  AND Duration_of_rheumatic_diseases < 3
  AND Number_of_painful_joints < 6
  AND Number_of_swollen_joints < 1
  AND Duration_of_morning_stiffness < 0.1
  AND Mucosal_manifestations = no
  AND Eye_manifestations = no
THEN  Diagnosis = Extraarticular_rheumatism [4 0 1 0 12 0 2 4]

IF    Sex = male
  AND 28 < Age < 74
  AND Joint_pain = arthritic
  AND Number_of_painful_joints < 11
  AND Spinal_pain = no
  AND Duration_of_morning_stiffness < 1.5
  AND Eye_manifestations = no
THEN  Diagnosis = Crystal_induced_synovitis [0 1 0 1 0 12 1 0]
```

Table 3
(continued)

```
IF    18 < Age < 33
  AND Family_anamnesis = no
  AND Duration_of_present_symptoms > 1
  AND Duration_of_rheumatic_diseases < 12
  AND Spinal_pain = no
  AND Other_manifestations = no
THEN  Diagnosis = Nonspecific_rheumatic_manifestations [4 1 0 3 4 1 17 5]

IF    Age < 62
  AND Joint_pain = no
  AND Number_of_painful_joints < 1
  AND Number_of_swollen_joints < 1
  AND Spinal_pain = no
THEN  Diagnosis = Nonrheumatic_diseases [6 3 0 1 5 1 5 19]
```

The induced diagnostic rules have an explicit symbolic form and can be understood and interpreted by specialists. The rules induced by CN2 with the Laplace estimate were shown to a specialist for rheumatic diseases. The analysis of rules by the expert has shown that most rules were consistent with the expert knowledge, although not very characteristic for specific diagnoses [20].

4.3. Performance evaluation on unseen cases

Since the ultimate test of the quality of induced rules is their performance on unseen examples, the second group of experiments was performed on ten different random partitions of the data set into 70% training and 30% testing examples. These are the same partitions as used by Lavrač et al. [20], from where the results of CN2 with the Laplace estimate are taken.

4.3.1. Results of rule induction

When using the m-estimate within CN2, the optimal value of m, chosen on each of the training partitions, was 32 for five partitions, 64 for four partitions and 128 for one partition. Judging by the high optimal values for the parameter m the data contains a relatively high degree of noise. This agrees with the opinion of the medical expert and earlier experiences with applying machine learning approaches in this domain.

The relative information scores of the induced rules are given in Table 4, while the classification accuracies are given in Table 5. Both tables display performance results on unseen cases. Table 5 also contains performance results for the k-NN algorithm, discussed in the following subsection.

To compare the results of CN2 with the m-estimate and CN2 with the Laplace estimate, we used the two-tailed statistical t-test for dependent samples. The six percentage points difference in the relative information scores is significant at the 99.99% level. The one percentage point difference in the accuracy is not significant even at the 50% level. We can conclude that CN2 with the m-estimate performs clearly better, as it achieves a much higher relative information score and an insignificantly lower classification accuracy.

4.3.2. Results of instance-based learning

Given the whole training set of 462 instances, k-NN stores all of them, without performing any generalization. Its classification accuracy on the training set is thus 100%, regardless of the use of feature weights. The weights assigned to the features are shown in Table 6. Six attributes are

Table 4
Relative information scores of rules derived by CN2 with the m-estimate and the Laplace estimate, measured on the testing set for each of the ten partitions

Partition	CN2 m-estimate (%)	CN2 Laplace estimate (%)
1	24	17
2	27	20
3	23	17
4	27	17
5	21	21
6	27	15
7	28	21
8	22	21
9	27	16
10	28	23
Average	25	19

Table 5
Classification accuracy of rules derived by CN2 with the m-estimate and the Laplace estimate, as well as classification accuracy of the k-NN algorithm without and with feature weights, measured on the testing set for each of the ten partitions

Partition	CN2 m-estimate (%)	CN2 Laplace estimate (%)	k-NN no weights (%)	k-NN with weights (%)
1	45.3	47.5	46.8	47.5
2	41.7	45.3	51.1	46.0
3	43.9	51.1	48.9	51.1
4	49.6	44.6	46.0	46.8
5	33.8	46.0	45.3	43.9
6	46.8	49.6	53.2	54.0
7	41.7	44.6	45.3	51.8
8	42.4	41.0	44.6	49.0
9	47.5	43.9	48.2	54.7
10	48.2	39.6	42.5	47.5
Average	44.1	45.3	47.2	49.2

highly relevant to the diagnostic task at hand: these are the number of swollen joints, the duration of rheumatic diseases, the duration of morning stiffness, the duration of present symptoms, the number of painful joints, and age. Mucosal manifestations and family anamnesis seem to be irrelevant.

The performance of the k-NN algorithm on unseen cases is shown in the right half of Table 5. The best values of k ranged from 15 to 41, with an average of 23. This again indicates that the data are quite noisy.

k-NN performs better than CN2, even without feature weights. According to the two-tailed statistical t-test for dependent samples, the differences in performance between k-NN without feature weights and CN2 is significant at the 90% level when the m-estimate is used and at the 95% level when the Laplace estimate is used. For k-NN with feature weights, the significance levels are 99% and 98%, respectively. The two percentage points difference in performance between k-NN with and without feature weights is significant at the 88% level.

Table 6
Feature weights for the 16 anamnestic attributes in the domain of early diagnosis of rheumatic diseases

Weight	Attribute
2.959	Number of swollen joints
2.825	Duration of rheumatic diseases
2.777	Duration of morning stiffness
2.681	Duration of present symptoms
2.569	Number of painful joints
2.038	Age
0.772	Spinal pain type
0.381	Joint pain type
0.291	Sex
0.249	Other pain type
0.239	Other manifestations
0.236	Skin manifestations
0.177	Eye manifestations
0.123	Therapy
0.090	Mucosal manifestations
0.080	Family anamnesis

5. The utility of background knowledge

The available patient data may be augmented with additional diagnostic knowledge which can be considered as additional information by the learner. In machine learning terminology, additional expert knowledge is usually referred to as *background knowledge.*

5.1. Background knowledge about rheumatic diseases

A specialist for rheumatic diseases has provided his knowledge about the typical co-occurrences of symptoms. Six typical groupings of symptoms were suggested by the specialist as background knowledge to be considered by the learner [20].

The first grouping relates the attribute 'Joint pain' and the attribute 'Duration of morning stiffness'. The characteristic combinations are given in Table 7, all other combinations are insignificant or irrelevant.

The second grouping relates the spinal pain and the duration of morning stiffness. The following are the characteristic combinations: no spinal pain and morning stiffness up to 1 hour, spondylotic pain and morning stiffness up to 1 hour, spondylitic pain and morning stiffness longer than 1 hour.

The third grouping relates the attributes sex and other pain. Indicative is the pain in the thorax or in the heels for male patients, all other combinations are non-specific: the corresponding values of Grouping 3 are thus 'male and thorax' and 'male and heels'.

The fourth grouping relates joint pain and spinal pain. All co-occurrences are characteristic: 'no pain and spondylotic', 'arthrotic and no pain', 'no pain and spondylitic', 'arthritic and spondylitic', 'arthritic and no pain', 'no pain and no pain'.

The fifth grouping relates joint pain, spinal pain and the number of painful joints, with characteristic values: 'no pain and spondylotic and 0', 'arthrotic and no pain and $1 \leqslant$ joints $\leqslant 30$', 'no pain and

Table 7
Characteristic combinations of values for the attributes 'Joint pain' and 'Duration of morning stiffness', as defined by the function 'Grouping 1'

Joint pain	Morning stiffness	Grouping 1 value
No pain	$\leqslant$ 1 hour	No pain and dms $\leqslant$1 hour
Arthrotic	$\leqslant$ 1 hour	Arthrotic and dms $\leqslant$1 hour
Arthritic	$>$ 1 hour	Arthritic and dms $>$ 1 hour

spondylitic and 0', 'arthritic and spondylitic and $1 \leqslant$ joints $\leqslant 5$', 'arthritic and no pain and $1 \leqslant$ joints $\leqslant$ 30', 'no pain and no pain and 0'.

The last, sixth, grouping relates the number of swollen joints and the number of painful joints. The characteristic values for Grouping 6 are: '0 and 0', '0 and $1 \leqslant$ npj $\leqslant 30$', and '$1 \leqslant$ nsj $\leqslant 10$ and $0 \leqslant$ npj $\leqslant 30$'.

The background knowledge is encoded in the form of functions, introducing specific function values for each characteristic combination of symptoms. All the other combinations (which are not explicitly specified above) have the same function value *irrelevant*. The characteristic combinations of attribute values are given names which are mnemonic and understandable. Their names are 'artificial' (not used by specialists), but they represent meaningful co-occurrences of symptoms which have their role in expert diagnosis.

5.2. *Learning with LINUS*

The main idea in LINUS [18, 19] is to incorporate different attribute-value learning algorithms into an environment for inductive logic programming [9, 18, 28], which enables the effective use of specialist background knowledge in learning as well as the induction of relational descriptions. Several attribute-value learners have been used within LINUS: a decision-tree induction algorithm ASSISTANT [4], the rule-induction algorithm NEWGEM [26], the CN2 algorithm, and the k-NN implementation of Wettschereck [37].

In addition to training examples, LINUS is given background knowledge represented in the form of logical definitions of relations or functions, such as the functional definitions of the symptom groupings above. Using the background knowledge, LINUS generates attributes that are not present in the initially given attribute set and extends the training examples with these attributes. The transformed problem can be addressed by an attribute-value learner. If the latter generates if–then rules, these can be transformed back in the form of logic programs.

LINUS thus extends attribute-value learners with the ability to use background knowledge and learn relational descriptions. On the other hand, many inductive logic programming (ILP) systems lack the ability to handle noisy data and real numbers, while many attribute-value learners have sophisticated mechanisms developed for that purpose. LINUS thus brings the possibility to use a variety of well-developed learning mechanisms within ILP.

We applied LINUS to the domain of early diagnosis of rheumatic diseases as described in Section 4 and the background knowledge described above. This yields six additional attributes, one for each of the six groupings of symptoms. The original examples are extended with the values of the new attributes. The new learning problem thus has 22 attributes and 462 examples.

5.3. Results of rule induction from the entire data set

The experiments we conducted on the extended learning problem were analogous to those performed on the basic problem. We first used CN2 with the Laplace- and m-estimate to induce rules from the whole example set. With the Laplace-estimate, CN2 induced a set of 38 rules (120 conditions), with an accuracy of 52.4% and a relative information score of 30% [20]. With the m-estimate, $m = 32$ proved to be the best value, yielding a set of 29 rules (121 conditions) with an accuracy of 64.5% and a relative information score of 46%. As compared to the results without background knowledge, a substantial increase in relative information score (8 percentage points) occurs when using the Laplace-estimate and slight improvement (1 percentage point) otherwise.

A selection of rules induced with CN2 using the m-estimate are given in Table 8. One rule per class was selected, and each of the selected rules covers at least five examples in addition to using one of the six symptom groupings provided by the specialist.

The weights used by k-NN for the six new attributes are given in Table 9. For the original attributes, the weights given in Table 6 were used. The relative importance of the groupings as determined by the weights corresponds to the number of their appearances in the induced rules: in the set of rules induced by CN2 with the Laplace-estimate, Grouping 4 appears nine times, Grouping 5 seven times, Grouping 2 three times, Grouping 1 twice, and Groupings 6 and 3 once each.

5.4. Performance evaluation on unseen cases

We also evaluated the performance of CN2 with the Laplace- and the m-estimate, as well as the performance of the k-NN algorithm, on the ten different partitions of the data set into 70% training and 30% testing examples.

The relative information scores of the rules induced by CN2 are given in Table 10: CN2 with the m-estimate performs slightly better, but the difference in performance is significant at the 80% level. Background knowledge improves the performance, especially for CN2 with the Laplace-estimate, but also for CN2 with the m-estimate. The differences are significant at the 99.9% and 80% levels, according to the two-tailed statistical t-test for dependent samples.

The classification accuracies of CN2 and of k-NN are given in Table 11. Background knowledge improves the performance in all cases. For CN2 with the m-estimate, CN2 with the Laplace-estimate, k-NN without feature weights and k-NN with feature weights, respectively, the differences are significant at the 90%, 98%, 88%, and 83% levels.

In terms of accuracy, CN2 with the Laplace-estimate performs better than CN2 with the m-estimate. However, it performs worse in terms of relative information score. k-NN with feature weights again performs better than CN2. The difference is significant at the 99.5% level for CN2 with the m-estimate and at the 98% for CN2 with the Laplace-estimate.

The best value of m for the ten partitions was 16 in four cases, 32 in three cases, and 64 in three cases. In nine of the ten cases, the best value of m was lower when background knowledge was given, i.e., the dataset appears to contain less noise when background knowledge is given. A similar effect can be noticed for the parameter k: it ranged from 7 to 21 with an average of 13 and was lower in the presence of background knowledge for eight of the ten partitions. This indicates that background knowledge alleviates the effects of data imperfections in this domain.

Table 8
A selection of rules for early diagnosis of rheumatic diseases that make use of specialist background knowledge. The rules were induced with CN2 using the m-estimate

```
IF    Age < 67
  AND Number_of_painful_joints < 2
  AND Other_pain = no
  AND Skin_manifestations = no
  AND grouping4(Joint_pain,Spinal_pain,no pain&spondylotic)
THEN  Diagnosis = Degenerative_spine_diseases [35 1 0 0 0 0 0 0]

IF    Sex = female
  AND Age > 46
  AND Number_of_painful_joints < 19
  AND grouping4(Joint_pain,Spinal_pain,arthrotic&no pain)
THEN  Diagnosis = Degenerative_joint_diseases [3 49 0 0 1 0 2 3]

IF    Duration_of_present_symptoms > 8
  AND Number_of_painful_joints < 3
  AND grouping5(Joint_pain,Spinal_pain,Number_of_painful_joints,
                arthrotic&no pain&1 =< joints =< 30)
THEN  Diagnosis = Other_inflammatory_diseases [0 0 0 5 0 2 0 1]

IF    Sex = male
  AND 28 < Age < 74
  AND Number_of_painful_joints < 17
  AND Number_of_swollen_joints < 8
  AND Eye_manifestations = no
  AND grouping3(Sex,Other_pain,irrelevant)
  AND grouping4(Joint_pain,Spinal_pain,arthritic&no pain)
THEN  Diagnosis = Crystal_induced_synovitis [0 1 0 1 0 12 0 0]

IF    Age < 44
  AND Duration_of_rheumatic_diseases < 10
  AND Number_of_painful_joints < 26
  AND Other_pain = no
  AND Skin_manifestations = no
  AND Other_manifestations = no
  AND grouping5(Joint_pain,Spinal_pain,Number_of_painful_joints,
                arthrotic&no pain&1 =< joints =< 30)
THEN  Diagnosis = Nonspecific_rheumatic_manifestations [2 2 0 1 1 0 8 3]

IF    33 < Age < 62
  AND Duration_of_present_symptoms > 2
  AND Number_of_swollen_joints < 1
  AND grouping5(Joint_pain,Spinal_pain,Number_of_painful_joints,
                no pain&no pain&0)
THEN  Diagnosis = Nonrheumatic_diseases [1 1 0 0 4 1 1 14]
```

6. Discussion

We have presented two machine learning approaches, rule induction and instance-based learning. The first constructs explicit symbolic (if–then) rules, which are generalizations of the training examples. Induced rules can then be used to classify new cases. The second stores the training examples and classifies new cases by comparing them to the stored cases.

In particular, we described the CN2 rule induction algorithm, which can use the Laplace- or the

Table 9
Feature weights for the six additional attributes derived from the domain knowledge of a specialist for rheumatic diseases

Weight	Attribute
1.086	Grouping 4
0.899	Grouping 5
0.698	Grouping 2
0.392	Grouping 1
0.345	Grouping 6
0.091	Grouping 3

Table 10
Relative information scores of rules induced by CN2 with the use of background knowledge, measured on the testing set for each of the ten partitions

Partition	CN2 m-estimate (%)	CN2 Laplace-estimate (%)
1	25	26
2	34	28
3	26	24
4	22	22
5	27	26
6	24	24
7	34	27
8	29	26
9	27	25
10	26	31
Average	27	26

Table 11
Classification accuracy in the presence of background knowledge for the CN2 and k-NN algorithms, measured on the testing set for each of the ten partitions

Partition	CN2 m-estimate (%)	CN2 Laplace-estimate (%)	k-NN no weights (%)	k-NN with weights (%)
1	45.3	48.9	48.9	52.5
2	47.5	51.8	49.6	53.2
3	49.6	48.9	52.5	51.8
4	46.8	48.2	46.8	47.5
5	46.0	46.8	48.9	49.6
6	43.2	48.2	47.5	50.4
7	49.6	48.9	52.5	55.4
8	48.9	48.2	48.2	46.8
9	48.9	48.2	51.1	54.0
10	46.0	48.2	45.3	48.2
Average	47.2	48.6	49.1	50.9

m-estimate to reliably estimate the classification accuracy of the rules considered in the induction process. The appropriate value of the parameter m is determined on the training set. We also described the k-NN algorithm, which can use feature weights to alleviate the negative influence of noisy or irrelevant features. The appropriate value of the parameter k is determined on the training set.

Both algorithms are capable of dealing with real data that involves real numbers and imperfections such as noise and missing values. They are thus suitable for applications in medical diagnosis and prognosis. We applied both approaches to the domain of early diagnosis of rheumatic diseases and compared their performance.

Let us first make some specific remarks about the methods used based on their performance on unseen cases in the domain studied. In CN2, the m-estimate yields better performance than the Laplace-estimate in terms of relative information score at approximately the same accuracy. Feature weights based on the mutual information between the features and the class improve the performance of k-NN. They also show the relative importance of attributes for classification. Finally, in this domain, k-NN performs better than CN2 in terms of classification accuracy.

Other studies [17] have also shown that k-NN in many cases performs better than symbolic learning approaches. However, a k-NN algorithm cannot explain its classifications as clearly as a rule induction system. It can provide as an explanation the k nearest neighbors used in the classification and their distances to the classified example, but this is more difficult to understand than a relatively short if–then rule, which explicitly refers to the attributes used. The rules can be also used without computer help and can become a part of the domain knowledge in the particular medical area, if accepted by specialists. To show that if–then rules are easy to understand, we have listed selected rules induced by CN2 with the m-estimate. A specialist has also judged the rules induced by CN2 with the Laplace-estimate as understandable and meaningful [20].

We have also briefly described the LINUS methodology for using specialist background knowledge in the learning process. We have applied this methodology to the examples and background knowledge in the domain of early diagnosis of rheumatic diseases, comparing the performance of CN2 and k-NN with and without background knowledge. The background knowledge improves performance in terms of relative information scores and classification accuracy, both for CN2 and for k-NN, k-NN still performing better than CN2.

In the presence of background knowledge, the best values for the parameters m and k substantially decreased. In the experiments with the ten partitions, the average best m was 35 with and 54 without background knowledge. Likewise, the average best k was 13 with and 23 without background knowledge. As higher values of m and k are appropriate for more noisy data, this indicates that the presence of background knowledge reduces the noise originally present in the data.

The classification accuracies achieved by both methods are around 50%. While this number is relatively low as compared to accuracies achieved in other domains, one should bear in mind that the patient records used contain only anamnestic data which is very unreliable and noisy by nature as well as the other problems with the data, described in Section 4.1. As the most common diagnosis of degenerative spine diseases accounts for only 34% of the patients, both methods succeed at extracting information from the patient records that is relevant for the diagnostic problem at hand. For the CN2 rules, this is confirmed by the positive relative information scores.

Our study indicates that rule induction and instance-based learning can be useful tools for medical diagnosis and prognosis. A combination of both approaches that gives both accurate predictions and satisfactory explanations may be the most appropriate approach, aimed at complementing the expertise of physicians with knowledge induced from stored patient records.

Acknowledgements

This research was financially supported by the Slovenian Ministry of Science and Technology and the ESPRIT Project 20237 Inductive Logic Programming II. The authors are grateful to the specialists of the University Medical Center in Ljubljana who helped collecting the data, especially to Vladimir Pirnat. Aram Karalič and Igor Kononenko prepared the data in a form appropriate for the experiments. At the time of writing this paper Sašo Džeroski was an ERCIM Fellow at the German National Research Center for Information Technology (GMD), Sankt Augustin, Germany.

References

[1] D. Aha, D. Kibler and M. Albert, Instance-based learning algorithms, *Machine Learning* **6** (1991), 37–66.
[2] I. Bratko and I. Kononenko, Learning rules from incomplete and noisy data, in: *Interactions in Artificial Intelligence and Statistical Methods*, B. Phelps, ed., Technical Press, Hampshire, 1987.
[3] B. Cestnik, Estimating probabilities: A crucial task in machine learning, in: *Proc. Ninth European Conference on Artificial Intelligence*, Pitman, London, 1990, pp. 147–149.
[4] B. Cestnik, I. Kononenko and I. Bratko, ASSISTANT 86: A knowledge elicitation tool for sophisticated users, in: *Progress in Machine Learning*, I. Bratko and N. Lavrač, eds, Sigma Press, Wilmslow, 1987, pp. 31–45.
[5] P. Clark and R. Boswell, Rule induction with CN2: Some recent improvements, in: *Proc. Fifth European Working Session on Learning*, Springer, Berlin, 1991, pp. 151–163.
[6] P. Clark and T. Niblett, The CN2 induction algorithm, *Machine Learning* **3**(4) (1989), 261–283.
[7] T.M. Cover and P.E. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **13** (1968), 21–27.
[8] B.V. Dasarathy, ed., *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques,* IEEE Computer Society Press, Los Alamitos, CA, 1990.
[9] L. De Raedt, ed., *Advances in Inductive Logic Programming*, IOS Press, Amsterdam, 1996.
[10] S. Džeroski, B. Cestnik and I. Petrovski, Using the m-estimate in rule induction, *Journal of Computing and Information Technology* **1** (1993), 37–46.
[11] S.A. Dudani, The distance-weighted k-nearest neighbor rule, *IEEE Transactions on Systems, Man, and Cybernetics* **6**(4) (1975), 325–327.
[12] E. Fix and J.L. Hodges, Discriminatory analysis. Nonparametric discrimination. Consistency properties. Technical Report 4, US Air Force School of Aviation Medicine, Randolph Field, TX, 1957.
[13] S. French, *Decision Theory*, Ellis Horwood, Chichester, 1986.
[14] R. Holte, L. Acker and B. Porter, Concept learning and the problem of small disjuncts, in: *Proc. Tenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA, 1989.
[15] A. Karalič and V. Pirnat, Machine learning in rheumatology, *Sistemica* **1**(2) (1990), 113–123.
[16] I. Kononenko and I. Bratko, Information-based evaluation criterion for classifier's performance, *Machine Learning* **6**(1) (1991), 67–80.
[17] I. Kononenko and M. Kukar, Machine learning for medical diagnosis, in: *Proc. Workshop on Computer-Aided Data Analysis in Medicine, CADAM-95*, IJS Scientific Publishing, Ljubljana, 1995.
[18] N. Lavrač and S. Džeroski, *Inductive Logic Programming: Techniques and Applications,* Ellis Horwood, Chichester, 1994.
[19] N. Lavrač, S. Džeroski and M. Grobelnik, Learning nonrecursive definitions of relations with LINUS, in: *Proc. Fifth European Working Session on Learning*, Springer, Berlin, 1991, pp. 265–281.
[20] N. Lavrač, S. Džeroski, V. Pirnat and V. Križman, The utility of background knowledge in learning medical diagnostic rules, *Applied Artificial Intelligence* **7** (1993), 273–293.
[21] L. Lesmo, L. Saitta and P. Torasso, Learning of fuzzy production rules for medical diagnosis, in: *Approximate Reasoning in Decision Analysis*, M. Gupta and E. Sanchez, eds, North-Holland, Amsterdam, 1982.
[22] P.J.F. Lucas, Logic engineering in medicine, *The Knowledge Engineering Review* **10**(2) (1995), 153–179. Cambridge University Press.
[23] R.S. Michalski, J.G. Carbonell and T.M. Mitchell, eds, *Machine Learning: An Artificial Intelligence Approach*, Vol. I, Tioga, Palo Alto, CA, 1983.
[24] R.S. Michalski, I. Mozetič, J. Hong and N. Lavrač, 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, San Mateo, CA, 1986, pp. 1041–1045.

[25] D. Michie, D.J. Spiegelhalter and C.C. Taylor, eds, *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, Chichester, 1994.
[26] I. Mozetič, NEWGEM: Program for learning from examples. Technical documentation and user's guide. Reports of Intelligent Systems Group UIUCDCS-F-85-949, Department of Computer Science, University of Illinois, Urbana Champaign, IL, 1985.
[27] S. Muggleton, *Inductive Acquisition of Expert Knowledge*, Addison-Wesley, Wokingham, 1990.
[28] S. Muggleton, ed., *Inductive Logic Programming*, Academic Press, London, 1992.
[29] M. Nunez, Decision tree induction using domain knowledge, in: *Current Trends in Knowledge Acquisition*, B. Wielinga, ed., IOS Press, Amsterdam, 1990.
[30] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA, 1988.
[31] V. Pirnat, I. Kononenko, T. Janc and I. Bratko, Medical analysis of automatically induced rules, in: *Proc. 2nd European Conference on Artificial Intelligence in Medicine*, Springer, Berlin, 1989, pp. 24–36.
[32] J.R. Quinlan, Induction of decision trees, *Machine Learning* **1**(1) (1986), 81–106.
[33] D.E. Rumelhart and J.L. McClelland, eds, *Parallel Distributed Processing, Vol. 1: Foundations*, MIT Press, Cambridge, MA, 1986.
[34] C.E. Shannon, A mathematical theory of communication, *Bell. Syst. Techn. J.* **27** (1948), 379–423.
[35] S.M. Weiss and C.A. Kulikowski, *Computer Systems that Learn*, Morgan Kaufmann, San Mateo, CA, 1991.
[36] D. Wolpert, Constructing a generalizer superior to NETtalk via mathematical theory of generalization, *Neural Networks* **3** (1989), 445–452.
[37] D. Wettschereck, A study of distance-based machine learning algorithms. PhD Thesis, Department of Computer Science, Oregon State University, Corvallis, OR, 1994.