Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries

Igor Zelič
INFONET
Planina 3
4000 Kranj
Slovenia

Igor Kononenko
Faculty of Computer and
Information Science
Tržaška 25
1001 Ljubljana
Slovenia

Nada Lavrač
J. Stefan Institute
Jamova 39
1001 Ljubljana
Slovenia

Vanja Vuga
University Medical Centre
Center for Sport Medicine
Celovška 25
1001 Ljubljana
Slovenia

Abstract

Machine learning techniques can be used to extract knowledge from data stored in medical databases. In our application, various machine learning algorithms were used to extract diagnostic knowledge to support the diagnosis of sport injuries. The applied methods include variants of the Assistant algorithm for top-down induction of decision trees, and variants of the Bayesian classifier. The available dataset was insufficent for reliable diagnosis of all sport injuries considered by the system. Consequently, expert-defined diagnostic rules were added and used as pre-classifiers or as generators of additional training instances for injuries with few training examples. Experimental results show that the classification accuracy and the explanation capability of the naive Bayesian classifier with the fuzzy discretization of numerical attributes was superior to other methods and was estimated as the most appropriate for practical use.

1 Introduction

Machine learning technology is well suited for the induction of diagnostic and prognostic rules and solving of small and specialized diagnostic and prognostic problems. Data about correct diagnoses/prognoses is often made available from archives of specialized hospitals and clinics, where the number of stored cases grows daily; similar data gathering is done also in daily routine of specialist medical doctors.

Such data gathering occurs also in the Center for Sport Medicine of the Ljubljana University Medical Hospital, where records of patients with sport injuries are collected daily. This work is limited to data analysis of patient records with injuries in athletics and handball. The reason for this limitation is a large number of possible diagnoses. Moreover, even in this limited domain, diagnoses have to be merged into diagnostic classes, in order to provide for a reasonable number-of-patients vs. number-of-diagnoses proportion needed for a successful application of machine learning methods.

The aim of this work is to provide systematic computer-supported data gathering and storing, intelligent analysis of stored data, support of diagnostic decisions, and the transfer of expert diagnostic knowledge from the experienced specialist to young inexperienced medical doctors. An important aspect, which has motivated this study, is to reveal the unclear influence of individual anamnestic and clinical parameters for individual diagnoses. Moreover, to support diagnostic decisions, reasonably high diagnostic accuracy has to be achieved, as well as the transparency of proposed solutions.

In recent years, many different machine learning systems were developed. Machine learning methods [16; 18] can be classified into three major groups [13]: inductive learning of symbolic rules (such as induction of rules [17; 6], decision trees [20] and induction of logic programs [15]), statistical or pattern-recognition methods (such as k-nearest neighbors or instance-based learning [7; 1], discriminate analysis and Bayesian classifiers), and artificial neural networks (such as networks with back-propagation learning, Kohonen's self-organizing network and Hopfield's associative memory [2]). In this work we are biased towards systems that provide for the explanation of proposed decisions, therefore the application of the 'black-box' neural networks was considered inappropriate; we have also limited the selection of systems to several variants of top-down decision tree learners and to several variants of the Bayesian classifier that have proved to be well suited for supporting diagnostic decision making in numerous medical domains [10].

The paper is organized as follows. Section 2 briefly describes the algorithms used in our study. Section 3 gives the description of the problem of sport injury diagnosis. Experiments and results are described in Section 4 together with the evaluation of the results by a medical expert. The paper concludes by discussing the advan-

tages and disadvantages of applied systems, and gives the directions for further work.

2 Machine learning systems

In this study we used only systems that provide for the explanation of decisions. We used several decision tree learners and several variants of the Bayesian classifier.

2.1 Decision tree learners

Three variants of the Assistant algorithm were used in our experiments: Assistant-R, Assistant-I, and Assistant-R2 [14].

Assistant-R is a reimplementation of the Assistant learning system for top down induction of decision trees [4]. The basic algorithm goes back to CLS (Concept Learning System) developed by Hunt et al. [8] and reimplemented by several authors (see [20] for an overview). The main features of Assistant are the binarization of attributes, decision tree prepruning and postpruning, incomplete data handling, and the use of the naive Bayesian classifier to calculate the classification in 'null leaves' (leaves for which no evidence is available, i.e., no training example falls into a null leaf).

The main difference between Assistant and its reimplementation Assistant-R is that ReliefF is used as a heuristic for attribute selection [11]. ReliefF is an extended version of RELIEF, developed by Kira and Rendell [9], 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. For example, RELIEF can efficiently estimate the quality of attributes in parity problems. In addition, wherever appropriate, instead of the relative frequency, Assistant-R uses the m-estimate of probabilities, which was shown to often improve the performance of machine learning algorithms [3].

Assistant-I is a variant of Assistant-R that, instead of ReliefF, uses the information gain as the selection criterion, the same as the original Assistant algorithm. However, other differences to Assistant remain (such as the use of the m-estimate of probabilities).

Assistant-R2 is a variant of Assistant-R that, instead of building one general decision tree for the whole domain, generates one decision tree for each class (diagnosis). When classifying a new instance all trees are tried. If several trees classify the instance into its coresponding class the most probable class is selected. If none of the trees 'fires' the general tree for all the diagnoses generated by Assistant-R is used.

2.2 Bayesian classifiers

Two variants of the Bayesian classifier are described, as well as their mechanisms for handling continuous attributes.

Naive Bayesian classifier uses the naive Bayesian formula to calculate the probability of each class C given the values V_i of all the attributes for a given instance to be classified, assuming the conditional independence of the attributes given the class:

$$P(C|V_1..V_n) = P(C)\prod_i \frac{P(C|V_i)}{P(C)} \tag{1}$$

A new instance is classified into the class with the maximal probability. We use the m-estimate [3] for computing the estimate of conditional probabilities:

$$P(C|V_i) = \frac{N(C\&V_i) + m \times P(C)}{N(V_i) + m} =$$

$$\frac{N(C\&V_i)}{N(V_i) + m} + \frac{m \times P(C)}{N(V_i) + m} \tag{2}$$

where $N(Cond)$ stands for the number of examples for which $Cond$ is fulfilled, and m is a user-defined parameter. The parameter m trades-off the contribution of the relative frequency and the prior probability. In our experiments, the parameter m was set to 2.0 (this setting is usually used as a default and, empirically, gives satisfactory results [3]).

For computing the prior probability, the Laplace law of succession is used [19]:

$$P(C) = \frac{N(C) + 1}{N_{ex} + N_{cl}} \tag{3}$$

where N_{ex} stands for the number of examples and N_{cl} for the number of classes.

The relative performance of the naive Bayesian classifier can serve as an estimate of the conditional independence of attributes.

Semi-naive Bayesian classifier is an extension of the naive Bayesian classifier that explicitly searches for dependencies between the values of different attributes [10]. If such a dependency is discovered, then the two values V_i and V_j of two different attributes are not considered as conditionally independent but rather as dependent by replacing in equation (1) the term

$$\frac{P(C|V_i)}{P(C)} \times \frac{P(C|V_j)}{P(C)}$$

by

$$\frac{P(C|V_i, V_j)}{P(C)}$$

For such a replacement, the reliable approximation of the conditional probability $P(C|V_i, V_j)$ is required. Therefore, the algorithm trades-off the non-naivety and the reliability of the approximations of probabilities.

Continuous attributes have to be prediscretized in order to be used by the (semi)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 that attribute. Discretization can be done manually by a domain expert or by applying a discretization algorithm [21].

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 overcomes this problem by considering the values of the continuous attribute (or, equivalently, the boundaries of intervals) as fuzzy values instead of point values [10]. 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.

2.3 Explanation capability of systems

In medical diagnosis it is crucial that the system is able to explain and argue about its decisions when diagnosing a new patient. Especially when faced with an unexpected solution of a new problem, the user requires substantial argumentation and explanation.

Decision tree learners are known to often give an appropriate explanation: induced decision trees are fairly easy to understand and can be used to support diagnosing without using the computer - this is particularly valuable in situations which require prompt decisions and in situations when computer interaction is psychologically inacceptable. Positions of attributes in the tree, especially the top (most informative) ones, often directly correspond to domain expert's knowledge. However, in order to produce general rules, these methods use pruning [20; 4] which drastically reduces tree sizes. Consequently, the paths from the root to the leaves are shorter, containing only few most informative attributes. Frequently physicians dislike such trees since too few parameters are taken into account and the tree too poorly describes the patients to provide for reliable decisions. Another problem is the variability of decision trees - frequently a small change in the dataset causes a substantial restructuring of the decision trees - this also decreases the physician's trust in the proposed diagnosis and in its explanation.

Bayesian classifiers induce a table of conditional probabilities which indicate how much a feature (an attribute value) contributes to a diagnosis. When explaining a decision for an individual patient, the explanation of a decision is provided by the indicated 'weigth' of a feature, i.e., the information gain for each patient's feature, as well as the sum of information gains of all features that are in favour or against the decision (diagnosis). The information gain (measured in bits) is computed as

$$-log_2 P(C) + log_2 P(C|V_i)$$

where C denotes an individual diagnosis, V_i an individual feature (attribute value), $P(C)$ the prior probability, and $P(C|V_i)$ the conditional probability. One of the main advantages of this approach, which is appealing to physicians, is that all the available information is used to explain the decision; such an explanation seems to be 'natural' for medical diagnosis and prognosis.

3 The diagnostic problem

In the Center for Sport Medicine of the Ljubljana University Medical Hospital, records of patients with sport injuries are collected daily. A patient's first visit to the Center results in a diagnosis of the injury and a treatment recommendation. Patients are usually treated by a series of therapeutic measures in a number of consecutive visits to the Center, as well as with recommended home treatment and exercising.

The current database of athletic and handball injuries consists of 118 patient records, described by values of 49 attributes. Although it is clear that for various diagnoses attributes have a varied diagnostic importance, the expert can identify the most important diagnostic attributes which include, for instance, the localization of injury, the test of forced movement, and the Lahman's test. Diagnoses are grouped into 30 diagnostic classes (the original database deals with more than 50 diagnoses). The most frequent diagnosis is the injury of ligamentary insertions (16% of patients have this diagnosis), thus the majority class is not significantly larger than other classes, which have 11% (injury of muscles of the back side of the thigh), 10% (injury of ankle joint), etc. There are 11 diagnostic classes represented by a single training instance, and 4 classes with two training instances each. A reasonable grouping of similar diagnoses only partly solves the problem of classes with too few instances. For example, the merging of diagnoses 'distensions of muscles semitendinosus' and 'distension of biceps feminoris' into a common diagnostic group 'injury of muscles of the back side of the thigh' is justified by the same location of muscles (as they are both located at the back side of the thigh), the same symptoms, similar treatment, and the same physiological cause of the two injuries.

3.1 Insufficient number of training instances

For the given large number of diagnostic classes, the number of available training instances is much too small to provide for reliable diagnostic decisions. The reason for a small number of instances is that only recently systematic data gathering has started, after having defined and selected the relevant diagnostic attributes and their values. The expert could, by browsing through old patient records, provide more training examples. However, due to the time constraints in the expert physician's daily practice, the expert's preference is to add new patient records when dealing with new patients. Consequently, our decision support system provides also a user-friendly interface which allows for inputing of new training instances with type/range consistency checking, as well as possible modifications of domain characteristics (classes, attributes, values, ranges of values, ...). The expert for sport injuries already made a modification by adding two new classes to the same domain: prognosis and therapy.

3.2 Dealing with few training instances

The problem of diagnoses with an insufficient number of training instances was (temporarly) solved by providing for a combined expert system—machine learning interaction when classifying new diagnostic cases. The system supports the input of expert-defined rules that have the same form as the training examples themselves. However, the expert defines only those attribute values that are characteristic for the diagnosis, whereas unimportant attributes remain undefined (value 'unknown'). The rule acquisition system checks the consistency of expert-defined rules with respect to the database of patient records. The rules are supposed to cover the stored instances of a given diagnostic class, and not to cover instances of other diagnostic classes. Discovered inconsistencies are reported and inconsistency elimination is recommended.

Expert-defined diagnostic rules can be used in two ways: as pre-classifiers or as generators of additional training instances.

- In pre-classification mode, rules are 'fired' in preprocessing, before using a machine learning classification mechanism. Thus, test examples, covered by one of the rules, are classified before machine learning classification starts. The expert system enables us to choose to either include or exclude the training examples, covered by one of the rules. We tested both options.
- In example-generation mode, rules can be used to 'artificially' generate new training examples. For each rule the system generates n artificial training examples (n is user defined). Each of artifical training examples is generated as follows: for each atribute whose value is defined in the expert's rule, set the value; for each attribute whose value is not defined in the expert's rule, set its value to 'unknown'. In this mode, by providing for expert-defined rules, the physician helps the system to increase the number of training examples. However, one has to be aware that adding artificial training examples affects the distribution of training examples; therefore, parameter n should be reasonably small.

For brevity, in the next section only results of the pre-classification mode are given. The results of the example-generation mode are similar.

4 Experiments and results

Since the ultimate test of the quality of learners is their performance on unseen cases, experiments were performed on ten different random partitions of the data into 70% training and 30% testing examples. In this way, ten training sets E_i and ten testing sets T_i, $i \in [1..10]$ were generated. In addition, partitions followed the rule that the training set must contain at leasta half of all the examples of each class. In the experiments all the systems used the same training and testing sets.

Results of the experiments in terms of the classification accuracy and information score[1] are outlined in tables below.

4.1 Results of experiments using decision tree learners

Table 1 summarizes the results of the Assistant algorithms, using the following parameter setting: $m = 2$, prepruning = off, and postpruning = on. All three variants of Assistant achieve approximately the same accuracy and absolute information score (note that the accuracy and information score are computed for pruned trees). The comparison of decision trees reveals that Assistant-I selects substantially different attributes than the other two variants and also generates slightly smaller decision trees, which are in turn slightly less accurate.

4.2 Results of experiments using Bayesian classifiers

The use of fuzzy boundaries significantly improves the classification accuracy of the naive Bayesian classifier (see Table 2). Although the number of continuous attributes is relatively small, strict discretization of continuous attributes exaggerates the importance of those

[1] The formulas for computing the absolute and relative information score are given in the Appendix. Tables of results give the absolute information scores.

Classifier	Accuracy (%)		Inf. score		Leaves (#)
	$\overline{x}$	σ	$\overline{x}$	σ	
Assistant-I	58.2	5.8	2.19	0.28	20.9
Assistant-R	62.9	5.7	2.25	0.21	26.3
Assistant-R2	61.7	6.2	2.22	0.06	3.2

Table 1. The performance of the Assistant algorithms, all using the same parameter setting. The number of leaves for Assistant-R2 is the average over 30 trees.

Classifier	Accuracy (%)		Inf. score	
	$\overline{x}$	σ	$\overline{x}$	σ
naive Bayes - strict	59.4	4.9	1.83	0.15
naive Bayes - fuzzy	69.4	3.0	2.32	0.19
semi-naive Bayes - fuzzy	59.4	4.8	1.82	0.15

Table 2. The performance of the Bayesian classifiers with m=2.

attributes. In the physician's opinion, the continuous attributes are not very important for classification; the fuzzy discretization correctly lowers their influence which importantly increases the classification accuracy.

The use of the semi-naive Bayesian classifier turns out to be inappropriate for this domain. Joining of values of attributes causes the accuracy to drop. The result suggests that in this domain the attributes are relatively conditionaly independent.

4.3 Results of experiments using expert-defined rules

Expert-defined rules are consistent with the training examples, i.e., they cover only instances with the same diagnosis as it appears in the conclusion part of the rule. We compared the performance of such a combined classifier by either including or excluding the training examples covered by expert-defined rules (see Table 3). The exclusion of these training instances does not change the classification accuracy significantly; however, the generated decision trees are much smaller. This suggests that the exclusion of instances covered by expert-defined rules importantly simplifies the model for other diagnoses.

4.4 Physician's evaluation of results

The expert physician is satisfied with the classification accuracy achieved by the naive Bayesian classifier and estimates this accuracy as acceptable. Besides, he likes the explanation of the decisions provided by the naive Bayesian classifier: he considers the sum of information gains in favour/against a given diagnosis to be close to the way how physicians diagnose patients. He also prefers the naive Bayesian classifier since it uses all the available attributes for classification.

On the other hand, the decision trees are not considered to be transparent. In fact the expert physician feels that the number of attributes in the tree is too small and that the classification with a decision tree ignores significant information about the patient. The decision tree generated by Assistant-I is even estimated as non-logical while the decision trees of Assistant-R do replicate the expert physician's knowledge about the most important attributes and their logical relations.

A possible reason for a relatively poor explanatory capacity of generated decision trees lies in a large number of diagnoses. In problems with a large number of decision classes, it is advisable to build trees that distinguish a selected class (diagnosis) against all the other classes (a binary classification problem), thus generating decision trees which give a better characterisation of the selected diagnostic class, and consequently a better explanation of the proposed decisions.

4.5 Evaluation on an independent test set

To further evaluate the most promising classifiers, i.e. the naive Bayesian classifier and Assistant-R, we tested their performance and explanation ability on a completely independent set of 20 new patients that were recently treated in the Center for Sport Medicine in Ljubljana. Table 4 shows the classifications of three typical patients by the naive Bayesian classifier (NB) and by Assistant-R (A-R).

The naive Bayesian classifier achieved 70% of the classification accuracy, whereas in 85% of cases the correct diagnosis was one of the two most probable diagnoses proposed by the classifier. The accuracy of Assistant-R was worse: it achieved a 47% classification accuracy, and a 64% accuracy if two most probable predictions were considered.

The expert physician was pleased with good performance of the naive Bayesian classifier, which is obviously

Classifier	accuracy (%)		inf. score		leaves (#)	
	in.	ex.	in.	ex.	in.	ex.
naive Bayes - fuzzy	69.4	69.4	2.32	2.19	/	/
Assistant-I	58.2	57.6	2.19	2.13	20.9	13.0
Assistant-R	62.9	64.4	2.25	2.24	26.3	19.3
Assistant-R2	61.6	64.4	2.22	2.24	3.2	3.0

Table 3. The influence of in/ex-clusion of training instances covered by expert-defined rules. The number of leaves for Assistant-R2 is the average over 30 trees.

Pac. #	Diagnosis	OK	NB %	A-R %
1	muscle injuries	yes	92	4
1	tendinitis	no	5	0
1	joint-injuries (fingers)	no	0	14
1	contusions	no	0	13
1	syn.tractus ilitibialis	no	0	12
1	distensio	no	0	11
1	...	...	...	...
2	apophysitis tibiae	yes	57	26
2	ligamentary origin injury	no	34	24
2	tendinitis	no	8	0
2	muscle injuries	no	0	10
2	...	...	...	...
3	joint-injuries (fingers)	no	82	0
3	distorsio cubiti	yes	8	13
3	contusions	no	7	13
3	syn.tractus	no	0	12
3	distensio	no	0	11
3	...	...	...	...

Table 4: The predicted probabilities (%) of diagnoses for three selected new patients, given by the naive Bayesian classifier (NB) and Assistant-R (A-R); "OK = yes" denotes the correct diagnosis.

much less sensitive to the small number of training instances per each possible diagnosis than Assistant-R.

5 Discussion and further work

The classification accuracy of 70% achieved by the naive Bayesian classifier (achieved on 20 new cases, as well as on the 10 testing partitions of the dataset) is surprisingly high if we take into account that the number of different diagnoses is 30 and that only 118 training instances are available. The study suggests that our diagnostic problem is not very difficult due to the appropriate selection of attributes.

The naive Bayesian classifier uses all the available attributes, achieves the highest classification accuracy, and provides transparent explanation of its decisions. Therefore, it is considered as the most promising machine-learning classifier that can support physicians' decisions. Decision trees are considered to be inapropriate due to the low number of attributes that they take into account. These conclusions are in agreement with previous studies in medicine [10].

Assistant's performance was less successful, which is mostly due to the large number of diagnostic classes. In this domain, ReliefF used by Assistant-R is a better heuristic for estimating the quality of attributes than the standard information gain heuristic used by Assistant-I. ReliefF is a non-myopic heuristic that can correctly estimate the quality of highly dependent attributes. This advantageous feature leads the learning algorithm to discover domain regularities which are in accordance with the expert physician's knowledge. In further work we are planning to evaluate also the explanation capabilities of decision trees generated by Assistant-R2 which should reveal whether a large number of decision classes was one of the reasons for the expert's evaluation of generated decision trees as unsuitable for the explanation of decisions.

We expect that future collection of data will substantially increase the number of training instances. Physicians plan to input the data of new patients on-line. With new training instances we are expecting to achieve a better accuracy and roboustness of the system. We are also planning to extend the current expert system to include other sport injuries. The system will be used for supporting specialist's decisions as well as for educational purposes, i.e., to train medical students and non-specialist physicians.

Acknowledgements

This work was supported by the Slovenian Ministry of Science and Technology.

References

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

[2] Anderson, J.A. and Rosenfeld, E. (1988). *Neurocomputing: Foundations of Research*. The MIT Press.

[3] Cestnik B. (1990). Estimating probabilities: A crucial task in machine learning, *Proc. European Conf. on Artificial Intelligence*, Stockholm, August, 1990, pp. 147-149.

[4] Cestnik, B., Kononenko, I., and Bratko, I. (1987). ASSISTANT 86: A knowledge elicitation tool for sophisticated users. In I. Bratko and N. Lavrač, editors, *Progress in Machine Learning*, pages 31–45. Sigma Press, Wilmslow.

[5] Clark, P. and Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In *Proc. Fifth European Working Session on Learning*, pages 151–163. Springer, Berlin.

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

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

[8] Hunt, E., Martin, J., and Stone, P. (1966). *Experiments in Induction.* New York, Academic Press.

[9] Kira K. and Rendell L. (1992). A practical approach to feature selection. In: D.Sleeman and P.Edwards (eds.), *Proc. Int. Conf. on Machine Learning ICML-92* (Aberdeen, July 1992). Morgan Kaufmann, pp.249-256.

[10] Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7:317-337.

[11] Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In: F. Bergadano and L. de Readt (eds.), *Proc. European Conf. on Machine Learning ECML-94*, (Catania, Sicily, April 1994) , Springer Verlag, pp.171-182.

[12] Kononenko, I. and Bratko, I. (1991). Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6(1): 67–80.

[13] Kononenko, I., and Kukar, M. (1995). Machine learning for medical diagnosis. In: N. Lavrač (ed.) *Proc. Workshop on Computer-Aided Data Analysis in Medicine, CADAM-95*, (Bled, November 1995), IJS Scientific Publishing, Ljubljana.

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

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

[16] Michalski, R.S., Carbonell, J.G., and Mitchell, T.M., editors (1983). *Machine Learning: An Artificial Intelligence Approach*, Volume I. Tioga, Palo Alto, CA.

[17] Michalski, R.S., Mozetič, I., Hong, J., and Lavrač, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. In *Proc. Fifth National Conference on Artificial Intelligence*, pages 1041–1045. Morgan Kaufmann, San Mateo, CA.

[18] Michie, D., Spiegelhalter, D.J., and Taylor, C.C., editors (1994). *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, Chichester.

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

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

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

Appendix: The information score

The information score of induced rules [12] is a performance measure for classifiers. The most general answer a classifier can give is a probability distribution over the N_{cl} classes.

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

$$\begin{cases} -logP(C) + logP'(C), & P'(C) \geq P(C) \\ log(1-P(C)) - log(1-P'(C)), & P'(C) < P(C) \end{cases}$$

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 $P'(C) > P(C)$, negative if the answer is misleading $P'(C) < P(C)$, and zero if $P'(C) = P(C)$.

The *average information score* I_a of the answers of a classifier on a testing set consisting of examples $e_1, e_2, ..., e_t$ belonging to one of classes $C_1, C_2, ..., C_{N_{cl}}$ is calculated as:

$$I_a = \frac{1}{t} \times \sum_{k=1}^{t} I(e_k)$$