Fuzzy-Relational Classification: Combining Pairwise Decomposition Techniques with Fuzzy Preference Modeling

Eyke Hüllermeier and Klaus Brinker
Philipps-Universität Marburg
Department of Mathematics and Computer Science
{eyke,brinker}@informatik.uni-marburg.de

Abstract

This paper introduces a new approach to classification which combines pairwise decomposition techniques from machine learning with ideas and tools from fuzzy preference modeling. The approach, called fuzzy relational classification, effectively reduces the problem of classification to a problem of decision making based on a fuzzy preference relation. It will be shown that, by decomposing such a relation into a strict preference, an indifference, and an incomparability relation, it becomes possible to quantify different types of uncertainty in classification, and thereby to support more sophisticated classification and postprocessing strategies.

Keywords: Machine learning, classification, fuzzy preference relations, decision analysis.

1 Introduction

As one of the standard problems of supervised learning, the performance task of classification has been studied intensively in the field of machine learning. The arguably simplest type of classification problems are *dichotomous* (*binary*, *two-class*) problems for which a multitude of efficient and theoretically well-founded classification methods exists. Needless to say, however, practically relevant problems are rarely restricted to the binary case. One approach for tackling *polychotomous* problems is to use model classes that are able to represent a multi-class classifier, i.e., an $\mathcal{X} \to \mathcal{L}$ mapping for $|\mathcal{L}| > 2$, directly. An alternative strategy to approach such problems is to transform the original problem into several binary problems via a *class binarization* technique. The most popular class binarization technique is the unordered or one-against-rest binarization, where one takes each class in turn and learns a binary concept that discriminates this class from all other classes.

The key idea of the alternative *learning by pairwise comparison* (LPC) approach (aka pairwise classification, round robin learning, one-vs-one) is to transform an m-class problem into $m(m-1)/2$ binary problems, one for each pair of classes.[1] At classification time, a query instance is submitted to all binary models, and the predictions of these models are combined into an overall classification. In [5, 6], it was shown that pairwise classification is not only more accurate than the one-against-rest technique but that, despite the fact that the number of models that have to be learned is quadratic in the number of classes, pairwise classification is also more efficient (at least in the training phase) than one-against-rest classification.

This paper elaborates on another interesting aspect of the LPC approach: Assuming that every binary learner outputs a score in the unit interval (or, more generally, an ordered scale), and that this score can reasonably be interpreted as a “fuzzy preference” for the first in comparison with the second class, the complete ensemble of pairwise learners produces a *fuzzy preference relation.* The final classification decision is then made on the basis of this relation. In other words, the problem of classification has been reduced, in a first step, to a problem of decision making based on a fuzzy preference relation.

The novel aspect here is to look at the ensemble of predictions as a fuzzy preference relation. This perspective establishes a close connection between (pairwise) learning and fuzzy preference modeling, and therefore allows for applying techniques from the former field in the context of machine learning. In this paper, we are especially interested in exploiting techniques for decomposing a fuzzy (weak) preference relation into a preference structure consisting of a strict preference, an indifference, and an incomparability relation.

[1] Alternatively, one can consider a binary problem for every *ordered* pair of classes, in which case the total number of such problems is doubled. We shall come back to this point later on.

As will be argued in more detail later on, the latter two relations have a quite interesting interpretation and important meaning in the context of classification, where they represent two types of uncertainty: ambiguity and ignorance. Consequently, these relations can support more sophisticated classification strategies, including those that allow for partial reject options.

The remainder of the paper is organized as follows. Section 2 details the LPC approach to classification, and section 3 recalls the basics of fuzzy preference structures. The idea of classification based on fuzzy preference relations is outlined in section 4. Section 5 elaborates on an important element of this approach, namely learning weak preferences between class labels. First empirical results are presented in section 6, and section 7 concludes the paper.

2 Learning by Pairwise Comparison

As mentioned earlier, learning by pairwise comparison (LPC) transforms a multi-class classification problem, i.e., a problem involving $m > 2$ classes (labels) $\mathcal{L} = \{\lambda_1 \ldots \lambda_m\}$, into a number of *binary* problems. To this end, a separate model (base learner) $\mathcal{M}_{i,j}$ is trained for each *pair* of labels $(\lambda_i, \lambda_j) \in \mathcal{L}$. $\mathcal{M}_{i,j}$ is intended to separate the objects with label λ_i from those having label λ_j. If $(x, \lambda_a) \in \mathcal{X} \times \mathcal{L}$ is an original training example (revealing that instance x has label λ_a), then x is considered as a *positive* example for all learners $\mathcal{M}_{a,j}$ and as a *negative* example for the learners $\mathcal{M}_{j,a}$ $(j \neq a)$; those models $\mathcal{M}_{i,j}$ with $a \not\in \{i, j\}$ simply ignore this example.

At classification time, a query x is submitted to all learners, and each prediction $\mathcal{M}_{i,j}(x)$ is interpreted as a vote for a label. In particular, if $\mathcal{M}_{i,j}$ is a $\{0, 1\}$-valued classifier, $\mathcal{M}_{i,j}(x) = 1$ is counted as a vote for λ_i, while $\mathcal{M}_{i,j}(x) = 0$ would be considered as a vote for λ_j. Given these outputs, the simplest classification strategy is to predict the class label with the highest number of votes. A straightforward extension of the above voting scheme to the case of $[0, 1]$-valued (scoring) classifiers yields a *weighted voting procedure*: The score for label λ_i is computed by

$$r_i \overset{\mathrm{df}}{=} \sum_{1 \leq j \neq i \leq m} r_{i,j}, \tag{1}$$

where $r_{i,j} = \mathcal{M}_{i,j}(x)$, and again the label with the highest score is predicted.

The votes $r_{i,j}$ in (1) and, hence, the learners $\mathcal{M}_{i,j}$ are usually assumed to be (additively) *reciprocal*, that is,

$$r_{j,i} \equiv 1 - r_{i,j} \tag{2}$$

and correspondingly $\mathcal{M}_{i,j}(x) \equiv 1 - \mathcal{M}_{j,i}(x)$. Practically, this means that only one half of the $m(m-1)$ classifiers $\mathcal{M}_{i,j}$ needs to be trained, for example those for $i < j$. As will be explained in more detail below, this restriction is not very useful in our approach. Therefore, we will train the whole set of classifiers $\mathcal{M}_{i,j}$, $1 \leq i \neq j \leq m$, which means that no particular relation between $r_{i,j}$ and $r_{j,i}$ will be assumed.

3 Fuzzy Preference Structures

Considering the classification problem as a *decision problem*, namely a problem of deciding on a class label for a query input x, an output $r_{i,j} = \mathcal{M}_{i,j}(x)$ can be interpreted as a *preference* for label λ_i in comparison with label λ_j: The higher $r_{i,j}$, the more preferred is λ_i as a classification for x, i.e., the more likely λ_i appears in comparison with label λ_j. Correspondingly, the matrix

$$\mathcal{R} = \begin{bmatrix} - & r_{1,2} & \ldots & r_{1,m} \\ r_{2,1} & - & \ldots & r_{2,m} \\ \vdots & & & \vdots \\ r_{m,1} & r_{m,2} & \ldots & - \end{bmatrix} \tag{3}$$

obtained by collecting the outputs of the whole classifier ensemble can be interpreted as a *fuzzy* or *valued preference relation*. A classification decision can then be made on the basis of the relation (3). To this end, one can resort to corresponding techniques that have been developed and investigated quite thoroughly in fuzzy preference modeling and decision making [4]. In principle, the simple voting scheme (1) outlined in section 2 can be seen as a special case of such a decision making technique.

In this paper, our interest concerns the application of techniques for decomposing the relation $\mathcal{R}$ into three associated relations with different meaning. Suppose that $\mathcal{R}$ can be considered as a *weak preference relation*, which means that $r_{i,j} = \mathcal{R}(\lambda_i, \lambda_j)$ is interpreted as $\lambda_i \succeq \lambda_j$, that is, "label λ_i is at least as likely as label λ_j". From this relation, one can derive a *fuzzy preference structure* consisting of a *strict preference relation* $\mathcal{P}$, an *indifference relation* $\mathcal{I}$, and an *incomparability relation* $\mathcal{J}$. Referring to the class of t-norms [9] to operate on fuzzy preference degrees, a fuzzy preference structure can be defined as follows: Let (T, S, N) be a continuous De Morgan triplet consisting of a strong negation N, a t-norm T, and its N-dual t-conorm S; moreover, denote the T-intersection of two sets A and B by $A \cap_T B$ and the S-union by $A \cup_S B$. A fuzzy preference structure on $\mathcal{L}$ is a triplet $(\mathcal{P}, \mathcal{I}, \mathcal{J})$ of fuzzy relations satisfying

- $\mathcal{P}$ and $\mathcal{J}$ are irreflexive, $\mathcal{I}$ is reflexive;
- $\mathcal{P}$ is T-asymmetrical ($\mathcal{P} \cap_T \mathcal{P}^t = \emptyset$), $\mathcal{I}$ and $\mathcal{J}$ are symmetrical;

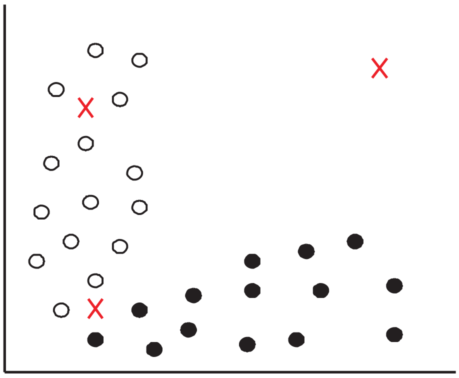

Figure 1: Classification scenario: Observations from two classes (points) and new query instances (crosses).

- $\mathcal{P} \cap_T \mathcal{I} = \emptyset$, $\mathcal{P} \cap_T \mathcal{J} = \emptyset$, $\mathcal{I} \cap_T \mathcal{J} = \emptyset$;
- $\mathcal{P} \cup_S \mathcal{P}^t \cup_S \mathcal{I} \cup_S \mathcal{J} = \mathcal{L} \times \mathcal{L}$.

The question of how to decompose a weak (valued) preference relation $\mathcal{R} \in [0,1]^{m \times m}$ into a strict preference relation $\mathcal{P}$, an indifference relation $\mathcal{I}$, and an incomparability relation $\mathcal{J}$ such that $(\mathcal{P}, \mathcal{I}, \mathcal{J})$ is a fuzzy preference structure have been studied extensively in the literature (e.g. [4, 1]). Without going into technical detail, we only give an example of a commonly employed decomposition scheme (again, we denote $r_{i,j} = \mathcal{R}(\lambda_i, \lambda_j)$):

$$\begin{aligned} \mathcal{P}(\lambda_i, \lambda_j) &= r_{i,j} \times (1 - r_{j,i}) \\ \mathcal{I}(\lambda_i, \lambda_j) &= r_{i,j} \times r_{j,i} \\ \mathcal{J}(\lambda_i, \lambda_j) &= (1 - r_{i,j}) \times (1 - r_{j,i}) \end{aligned} \tag{4}$$

A related decomposition scheme will also be used in the experimental part below.

4 Fuzzy Modeling of Classification Knowledge

The relations $\mathcal{I}$ and $\mathcal{J}$ have a very interesting meaning in the context of classification: Indifference corresponds to the *ambiguity* of a classification decision, while incompatibility reflects the corresponding degree of *ignorance*. To illustrate what we mean, respectively, by ambiguity and ignorance, consider the simple classification scenario shown in Fig. 1: Given observations from two classes, `black` and `white`, three new instances marked by a cross need to be classified. Obviously, given the current observations, the upper left instance can quite safely be classified as `white`. The case of the lower left instance, however, involves a high level of ambiguity, since both classes, `black` and `white`, appear plausible. The third situation is an example of ignorance: The upper right instance is located in a region of the instance space in which no observations have been made so far. Consequently, there is neither evidence in favor of class `black` nor in favor of class `white`.

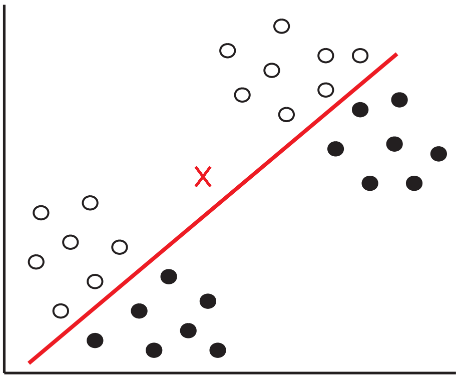

Figure 2: Given the assumption of linear separability, the query instance can be classified quite safely, even though it is spatially isolated from all other examples.

In the above example, the meaning of and difference between ambiguity and ignorance is intuitively quite obvious. Upon closer examination, however, these concepts turn out to be more intricate. In particular, one should realize that ignorance is not immediately linked with sparseness of the input space. This is due to the fact that generalization in machine learning is not only based on the observed data but also involves a model class with associated model assumptions. In fact, a direct connection between ignorance and sparely populated regions of the input space can only be established for instance-based (prototype-based) classifiers, since these classifiers are explicitly based on the assumption that closely neighbored instances belong to the same class.

The situation is different, however, for other types of models. For example, Fig. 2 shows a scenario in which a query point in a sparse input region can be classified quite safely, given the observed data *in conjunction with the assumption of a linear model*. In other words, given the correctness of the inductive bias of the learner (linearity assumption), the current observations allow for quite confident conclusions about the label of the query, even though the latter does not have any close neighbors.

The above considerations give rise to the following conception of ambiguity and ignorance in the context of classification: Let $\mathfrak{M}$ denote the model class underlying the classification problem, and let $\mathcal{V} = \mathcal{V}(\mathcal{D})$ be the set of models which are compatible with the examples given, i.e., the set of models which can still be regarded as possible candidates given the data $\mathcal{D}$; in the machine learning literature, $\mathcal{V}$ is called the *version* *space*. Now, given a query $x_0 \in \mathcal{X}$, the set of possible predictions is

$$Y_0 = \{\mathcal{M}(x) \,|\, \mathcal{M} \in \mathcal{V}(\mathcal{D}) \subseteq \mathfrak{M}\} \tag{5}$$

If the output of a model $\mathcal{M} \in \mathfrak{M}$ is a (deterministic) class label, then Y_0 is a subset of class labels ($Y_0 \subseteq \mathcal{L}$). Otherwise, if $\mathfrak{M}$ is a class of probabilistic classifiers, then Y_0 is a class of probability distributions over $\mathcal{L}$. In any case, it seems reasonable to define the degree of ignorance of a prediction in terms of the *diversity* of Y_0: The more predictions appear possible, i.e., the higher the diversity of predictions, the higher is the degree of ignorance.

According to this view, ignorance (incomparability) corresponds to that part of the (total) uncertainty about a prediction which can potentially be reduced by gathering more examples and thereby shrinking the version space. As opposed to this, the degree of ambiguity (indifference) corresponds to that part of the uncertainty which is due to a known conflict and which cannot be reduced any further.

The general idea of our method is to learn the weak preference relation (3), using an LPC approach, and to decompose this relation into a preference structure $(\mathcal{P}, \mathcal{I}, \mathcal{J})$ such that $\mathcal{J}$ characterizes the ignorance involved in a prediction, in the sense as outlined above, and $\mathcal{I}$ the ambiguity of the classification. In this context, two important questions have to be answered: Firstly, how to learn a suitable weak preference relation $\mathcal{R}$, and secondly, how to decompose $\mathcal{R}$ into a structure $(\mathcal{P}, \mathcal{I}, \mathcal{J})$. These problems will be discussed in more detail in the following section.

5 Learning Weak Preference Relations

As mentioned above, the first step of our method consists of learning the weak preference relation $\mathcal{R}$. More specifically, for every pair of labels (λ_i, λ_j), we have to induce models $\mathcal{M}_{i,j}$ and $\mathcal{M}_{j,i}$ such that, for a given query input x, $\mathcal{M}_{i,j}(x)$ corresponds to the degree of weak preference $\lambda_i \succeq \lambda_j$ and, vice versa, $\mathcal{M}_{j,i}(x)$ to the degree of weak preference $\lambda_j \succeq \lambda_i$.

The models $\mathcal{M}_{i,j}$ are of special importance as they directly determine the degrees of ambiguity and ignorance associated with a comparison between λ_i and λ_j. This fact is also crucial for the properties that the models $\mathcal{M}_{i,j}$ should obey.

According to the idea outlined in the previous section, a weak preference in favor of a class label should be derived from the set (5) of possible predictions. As this set in turn depends on the version space $\mathcal{V}$, the problem comes down to computing or at least approximating this space. In this connection, it deserves mentioning that an exact representation of the version space will usually not be possible for reasons of complexity. Apart from that, however, a representation of that kind would not be very useful either. In fact, despite the theoretical appeal of the version space concept, a considerable practical drawback concerns its extreme sensitivity toward noise and inconsistencies in the data.

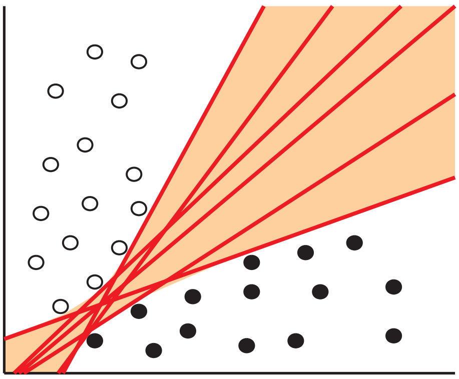

Figure 3: Illustration of the version space (class of hyperplanes that classify the training data correctly) and the "region of ignorance" (shaded in light color).

To overcome these problems, our idea is to approximate a version space in terms of a finite number of representative models. More specifically, consider the problem of learning a binary model $\mathcal{M}_{i,j}$ from an underlying model class $\mathfrak{M}$. To approximate the version space associated with $\mathcal{M}_{i,j}$, we induce a finite set of models

$$\mathbb{M}_{i,j} = \left\{ \mathcal{M}^1_{i,j}, \mathcal{M}^2_{i,j} \ldots \mathcal{M}^K_{i,j} \right\} \subseteq \mathfrak{M} \tag{6}$$

The set of possible predictions (5) is approximated correspondingly by

$$\widehat{Y}_0 = \mathbb{M}_{i,j}(x) = \bigcup_{k=1\ldots K} \mathcal{M}^k_{i,j}(x).$$

The way in which the models in (6) are obtained depends on the model class $\mathfrak{M}$. The basic idea is to apply randomization techniques as they are typically employed in ensemble learning methods. In the experiments below, we shall use ensembles of linear perceptrons, each of which is trained on a random permutation of the whole data.

An illustration is given in Fig. 3. Assuming that the two classes `black` and `white` can be separated in terms of a linear hyperplane, the version space consists of all those hyperplanes that classify the training data correctly. Given a new query instance, a unique class label can be assigned only if that instance lies on the same side of *all* hyperplanes (this situation is sometimes called "unanimous voting" [11]). Otherwise,

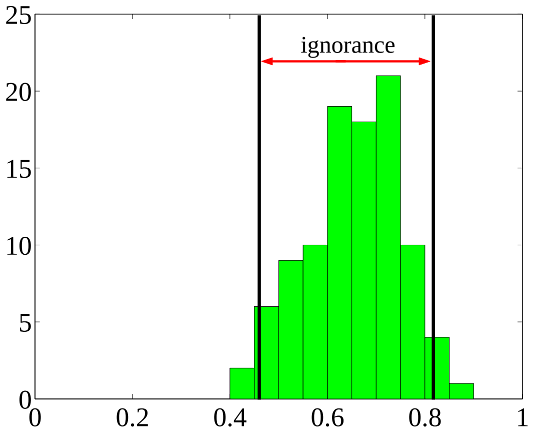

Figure 4: Distribution of the scores output by an ensemble $\mathbb{M}_{i,j}$. The degree of ignorance corresponds to the imprecision (width) of the distribution (here measured in a robust way in terms of the distance between the α- and $(1-\alpha)$-quantile).

both predictions are possible; the corresponding set of instances constitutes the "region of ignorance" which is shaded in light color.

In the above example, $\{0,1\}$-valued classifiers were used for the sake of simplicity. In the context of fuzzy classification, however, scoring classifiers with outputs in the unit interval are more reasonable. Suppose that each ensemble member $\mathcal{M}_{i,j}^k$ in (6) outputs a score $s_{i,j}^k \in [0,1]$. The minimum of these scores would in principle be suitable as a degree of (weak) preference for λ_i in comparison with λ_j:

$$r_{i,j} = \min_{k=1\ldots K} s_{i,j}^k .$$

As this order statistic is quite sensitive toward noise and outliers, however, we propose to replace it by the empirical α-quantile of the distribution of the $s_{i,j}^k$ (a reasonable choice is $\alpha = 0.1$).

Note that, in case the models in $\mathfrak{M}$ are reciprocal, only $\mathbb{M}_{i,j}$ or $\mathbb{M}_{j,i}$ needs to be trained, but not both. We then have $s_{i,j}^k = 1 - s_{j,i}^k$, and the α-quantile for $\mathcal{M}_{i,j}$ is given by 1 minus the $(1-\alpha)$-quantile for $\mathcal{M}_{j,i}$. In other words, the degree of ignorance is directly reflected by the distribution of the scores $s_{i,j}^k = 1 - s_{j,i}^k$ and corresponds to the length of the interval between the α-quantile and the $(1-\alpha)$-quantile of this distribution. Thus, the more precise this distribution, the smaller the degree of ignorance. In particular, if all models $\mathcal{M}_{i,j}^k$ output the same score s, the ignorance component shrinks to 0. An illustration is given in Fig. 4.

	name	# features	# examples
1	australian_scale	14	690
2	breast-cancer_scale	10	683
3	fourclass_scale	2	862
4	german	24	1000
5	heart_scale	13	270
6	splice_scale	60	1000
7	sonar_scale	60	208
8	w1a	300	2477

Figure 5: Data sets used in the experiments.

Our approach of *fuzzy relational classification* (FRC) as outlined above can be seen as a technique for deriving a condensed representation of the classification-relevant information contained in the version space. Once a preference structure $(\mathcal{P}, \mathcal{I}, \mathcal{J})$ has been induced, it can be taken as a point of departure for sophisticated decision strategies which go beyond simple voting procedures. This approach becomes especially interesting in extended classification scenarios, that is, generalizations of the conventional setting in which a single decision in favor of a unique class label is requested. For example, it might be allowed to predict several class labels instead of single one in cases of ambiguity, or to defer an immediate decision in cases of ignorance (or ambiguity). The latter scenario is known as *classification with reject option* in the literature, where one often distinguishes between *ambiguity rejection* [2, 7] and *distance rejection* [3]. Interestingly, this corresponds roughly to our distinction between ambiguity and ignorance. As we explained above, however, our conception of ignorance is more general and arguably more faithful, as it takes the underlying model assumptions into account: equating distance (between the query and observed examples) with ignorance does make sense for instance-based classifiers but not necessarily for other approaches with different model assumptions.

Of course, the design of suitable decision policies is highly application-specific and beyond the scope of this paper. In the next section, we therefore restrict ourselves to a simple experimental setup which is suitable for testing a key feature of FRL, namely its ability to represent the amount of uncertainty associated with a classification. More specifically, we used FRL as a means for implementing a reject option in the context of binary classification.

6 Experimental Results

We conducted an experimental study on 8 binary classification data sets from the Statlog and UCI repositories (cf. Fig. 5).[2] Each of the data sets was randomly

[2] These are preprocessed versions from the LIBSVM-website.

split into a training and test set of (roughly) equal size. As model classes $\mathbb{M}_{i,j}$, we used ensembles of 100 perceptrons with linear kernels and the default additive diagonal constant 1 (to account for non-separable problems), which were induced on the training data. Each perceptron was provided with a random permutation of the training set in order to obtain a diverse ensemble [8]. This process was repeated 10 times to reduce the bias induced by the random splitting procedure, and the results were averaged.

On the test sets, the real-valued classification outputs of the perceptrons were converted into normalized scores using a common logistic regression approach by Platt [10]. For a given test instance, the weak preference component $r_{i,j} = \mathcal{R}(\lambda_i, \lambda_j)$ was derived by the 0.1-quantile of the distribution of the scores from the ensemble $\mathbb{M}_{i,j}$ (see section 5). Moreover, as a decomposition scheme we used a slight modification of (4):

$$\begin{aligned}
\mathcal{P}(\lambda_i, \lambda_j) &= r_{i,j}\,(1 - r_{j,i}) \\
\mathcal{I}(\lambda_i, \lambda_j) &= 2\,r_{i,j}\,r_{j,i} \\
\mathcal{J}(\lambda_i, \lambda_j) &= 1 - (r_{i,j} + r_{j,i})
\end{aligned} \tag{7}$$

The reason for the modification is that in (7), the ignorance component nicely agrees with our derivation of weak preference degrees: It just corresponds to the width of the distribution of the scores generated by $\mathbb{M}_{i,j}$ (or, more precisely, the length of the interval between the quantiles of this distribution); therefore, it reflects the diversity of the predictions and becomes 0 if all ensemble members $\mathcal{M}_{i,j}^k$ agree on exactly the same score.

Finally, all test instances were ordered with respect to the associated degrees of indifference (ignorance), and corresponding accuracy-rejection diagrams were derived. These diagrams provide a visual representation of the accuracy levels α as a function of the rejection rate ρ: If the ρ% test instances with the highest degrees of indifference (ignorance) are refused, then the classification rate on the remaining test instances is α. Obviously, the effectiveness of FRL in representing uncertainty is in direct correspondence with the shape of the accuracy-rejection curve: If the degree of indifference (ignorance) produced by FRL is a good indicator of the reliability of a classification, then the ordering of instances according to indifference (ignorance) is in agreement with their respective degree of reliability (chance of misclassification), which in turn means that the accuracy-rejection curve is increasing. The presumption that FRL is indeed effective in this sense is perfectly confirmed by the experimental results, as can be seen in Fig. 6–7.

7 Conclusions

In this paper, we have introduced a new approach to classification learning which refers to the concept of fuzzy preference structures. This approach is intimately related with learning by pairwise comparison (LPC), a well-known machine learning technique for reducing multi-class to binary problems. The key idea of our approach, called *fuzzy relational classification* (FRC), is to use LPC in order to learn a fuzzy (weak) preference relation among the potential class labels. The original classification problem thus becomes a problem of decision making, namely of taking a course of action on the basis of this fuzzy preference relation. This way, our approach makes machine learning amenable to techniques and decision making strategies that have been studied intensively in the literature on fuzzy preferences.

An interesting example of corresponding techniques has been considered in more detail in this paper, namely the decomposition of a weak preference relation into a strict preference, an indifference, and an incomparability relation. We have argued that, in a classification context, indifference can be interpreted as the *ambiguity* of a prediction while indifference represents the level of *ignorance*. These concepts can be extremely useful, especially in extended classification scenarios which go beyond the prediction of a single label or do offer the option to abstain from a immediate classification decision.

First empirical studies have shown that FRC is indeed able to represent the uncertainty related to a classification decision: The implementation of a reject options turned out to be highly effective, regardless of whether the decision to abstain is made on the basis of the degree of ambiguity or the degree of ignorance.

The main contribution of this paper is a basic conceptual framework of fuzzy relational classification, including first empirical evidence in favor of its usefulness. Nevertheless, this framework is far from being complete and still leaves much scope for further developments. This concerns almost all steps of the approach and includes both aspects of learning and decision making. Just to give an example, our approach outlined in section 5 is of course not the only way to learn a weak preference relation. Moreover, the aspect of optimal decision making on the basis of pairwise preferences has not yet been addressed (as it strongly depends on the classification scenario). Issues of that kind will therefore be explored in future work.

Acknowledgements. This research was supported by the German Research Foundation (DFG).

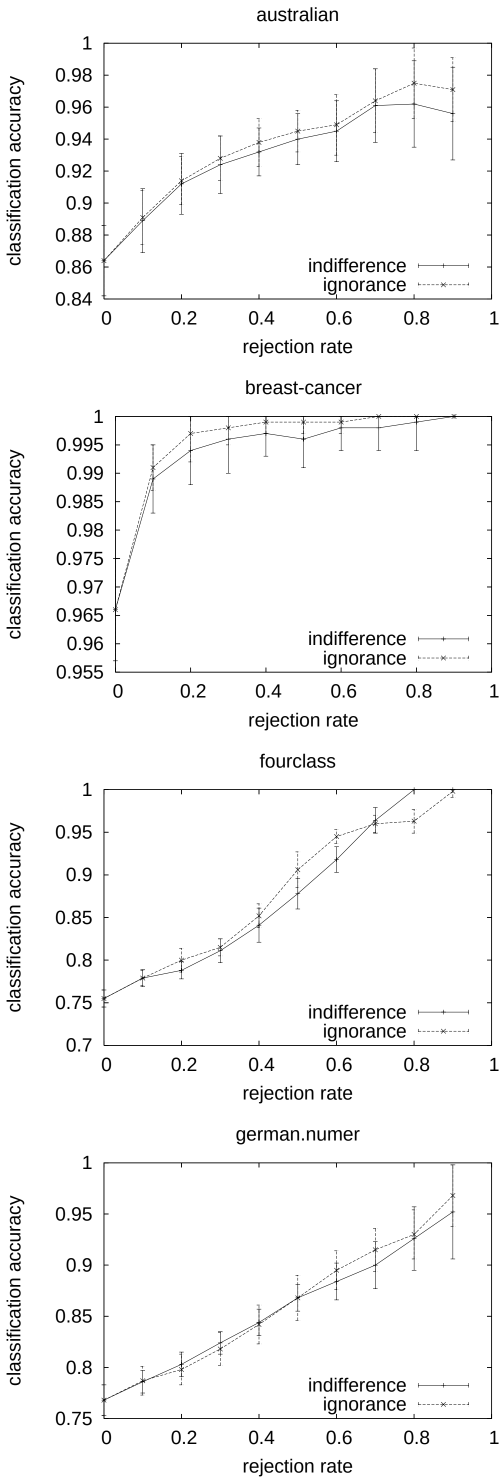

Figure 6: Accuracy-rejection curves for the data sets 1–4.

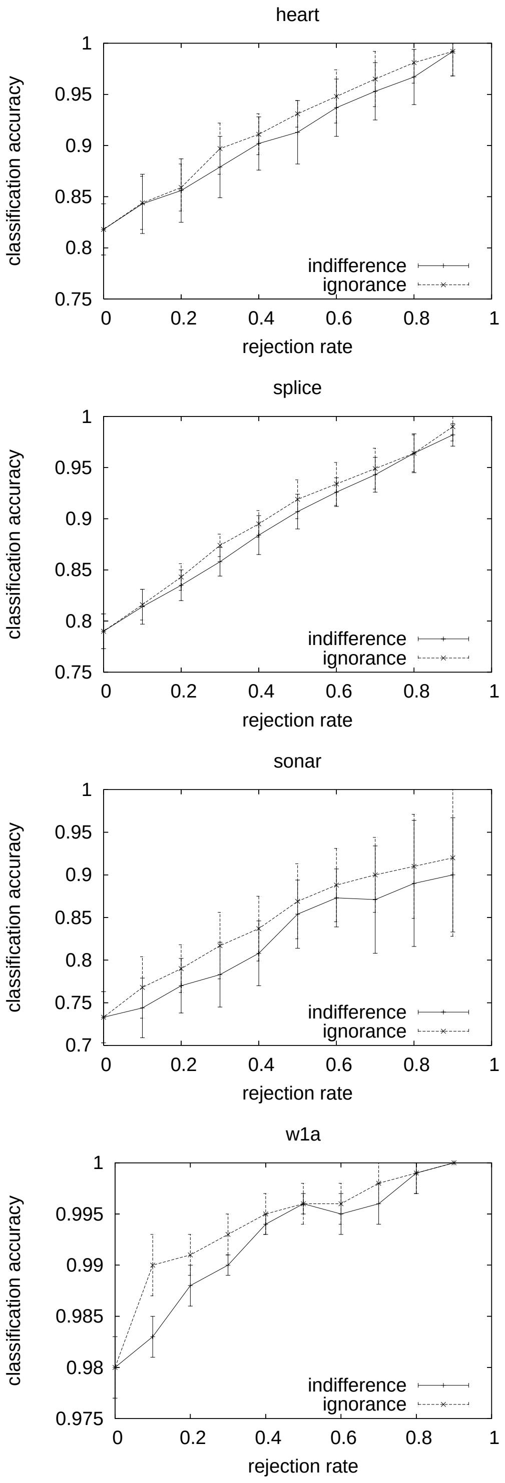

Figure 7: Accuracy-rejection curves for the data sets 5–8.

References

[1] B. De Baets, B. Van de Walle, and E. Kerre. Fuzzy preference structures and their characterization. *The Journal of Fuzzy Mathematics*, 3:373–381, 1995.

[2] CK. Chow. An optimum character recognition system using decision functions. *Trans. on Electronic Computers*, 6:247–253, 1957.

[3] B. Dubuisson and MH. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(1):155–165, 1993.

[4] J. Fodor and M. Roubens. *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer, 1994.

[5] J. Fürnkranz. Round robin rule learning. In *ICML-2001, Proc. 18th International Conference on Machine Learning*, pages 146–153, Williamstown, MA, 2001.

[6] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

[7] TM. Ha. The optimum class-selective decision rule. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(6):608–615, 1997.

[8] Ralf Herbrich. *Learning Kernel Classifiers*. MIT Press, 2002.

[9] EP. Klement, R. Mesiar, and E. Pap. *Triangular Norms*. Kluwer Academic Publishers, 2002.

[10] John Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A.J. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74, Cambridge, MA, 1999. MIT Press.

[11] E. Smirnov, I. Sprinkhuizen-Kuyper, G. Nalbantov, and S. Vanderlooy. Version space support vector machines. In *Proc. ECAI-06, 17th European Conference on Artificial Intelligence*, pages 809–810, Riva del Garda, Spain, 2006.