On Pairwise Naive Bayes Classifiers

Jan-Nikolas Sulzmann[1], Johannes Fürnkranz[1], and Eyke Hüllermeier[2]

[1] Department of Computer Science, TU Darmstadt
Hochschulstr. 10, D-64289 Darmstadt, Germany
{sulzmann,juffi}@ke.informatik.tu-darmstadt.de
[2] Informatics Institute, Marburg University
Hans-Meerwein-Str., Lahnberge, D-35032 Marburg, Germany
eyke@mathematik.uni-marburg.de

Abstract. Class binarizations are effective methods for improving weak learners by decomposing multi-class problems into several two-class problems. This paper analyzes how these methods can be applied to a Naive Bayes learner. The key result is that the pairwise variant of Naive Bayes is equivalent to a regular Naive Bayes. This result holds for several aggregation techniques for combining the predictions of the individual classifiers, including the commonly used voting and weighted voting techniques. On the other hand, Naive Bayes with one-against-all binarization is not equivalent to a regular Naive Bayes. Apart from the theoretical results themselves, the paper offers a discussion of their implications.

1 Introduction

The Naive Bayes classifier is a Bayesian learner that often outperforms more sophisticated learning methods such as neural networks, nearest neighbor estimation, or decision tree learning in many application areas. It is widely esteemed because of its simplicity, versatility, efficiency, and comprehensibility to domain experts (Kononenko, 1993). Even though the Naive Bayes classifier is directly amenable to multi-class problems, we consider the question whether its performance can be improved by combining it with class binarization methods. This question is motivated by the fact that class binarization has yielded good results for other multi-class learners as well (Fürnkranz, 2003).

The paper starts with a brief recapitulation of the Naive Bayes classifier (Section 2) and class binarization methods (Section 3). The main results are then presented in Section 4. We first derive a general method for combining pairwise Bayesian classifiers (Section 4.1), and then show that this method and the commonly used weighted and unweighted voting techniques are equivalent to the regular classifier (Sections 4.2 and 4.3). In Section 5, we address the same question for the alternative one-against-all class binarization technique and show that, in this case, equivalence to the regular Naive Bayes learner is lost. Finally, in Section 6, we briefly recapitulate the results and discuss some implications thereof.

J.N. Kok et al. (Eds.): ECML 2007, LNAI 4701, pp. 371–381, 2007.

2 Naive Bayes Classifier

Consider a simple setting of classification learning, in which the goal is to predict the class $c \in C = \{c_1, \ldots, c_m\}$ of an query input $x = (a_1, \ldots, a_n)$, given a training set of pre-classified examples; instances are characterized in terms of an attribute-value representation, and a_i is the value of the i^{th} attribute.

In general, the classification error can be minimized by selecting

$$\arg\max_{c_i \in C} \Pr(c_i|x), \tag{1}$$

i.e., the class with maximum posterior probability. To identify this class, estimates of the conditional probabilities $\Pr(c_i|x)$, $i = 1, \ldots, m$, are needed. Bayes theorem states that

$$p_i = \Pr(c_i|x) = \frac{\Pr(x|c_i) \cdot \Pr(c_i)}{\Pr(x)}, \tag{2}$$

and therefore allows one to reverse the original (direct) estimation problem into a more tractable (indirect) one: instead of estimating the probability of a class given the input, it suffices to estimate the probability of an input given a class.

The denominator in (2), $\Pr(x) = \sum_j \Pr(x|c_j) \cdot \Pr(c_j)$, is a normalizing constant that does not influence the solution of the maximization problem (1). Thus, the following basic version of a Bayesian learner is obtained:

$$\begin{aligned} c_B &= \arg\max_{c_i \in C} \Pr(x|c_i) \cdot \Pr(c_i) \\ &= \arg\max_{c_i \in C} \Pr(a_1, a_2, ..., a_n|c_i) \cdot \Pr(c_i) \end{aligned}$$

Under the so-called *Naive Bayes assumption*, which assumes the probabilities of attributes to be conditionally independent given the class, the difficult estimation of the (high-dimensional) probability $\Pr(x|c_i)$ can be reduced to the estimation of (one-dimensional) class-conditional attribute probabilities $\Pr(a_j|c_i)$:

$$\Pr(x|c_i) = \Pr(a_1, a_2, ..., a_n|c_i) \stackrel{!}{=} \prod_{j=1}^{n} \Pr(a_j|c_i)$$

The probabilities $\Pr(c_i)$ and $\Pr(a_j|c_i)$ can now be estimated from the training data, which is typically done by referring to corresponding relative frequencies. Even though the Naive Bayes assumption is usually violated in practice, and the probability estimates for $\Pr(c_i|x)$ are often not very accurate, the Naive Bayes prediction

$$c_{NB} = \arg\max_{c_i \in C} \left(\Pr(c_i) \cdot \prod_{j=1}^{n} \Pr(a_j|c_i) \right)$$

achieves surprisingly high classification rates. This is because, for a correct classification, it is only important that the true class receives the highest (estimated)

probability. In other words, only the *order* of the probability estimates $\Pr(c_i|x)$ is relevant, and this order is, to some extent, robust toward deviations of the estimated from the real probabilities (Domingos and Pazzani, 1997).

3 Class Binarization

Class binarization techniques turn multi-class problems into a set of binary problems. Prominent examples include one-against-all binarization (Clark and Boswell, 1991; Anand et al., 1995; Cortes and Vapnik, 1995; Rifkin and Klautau, 2004), pairwise classification (Friedman, 1996; Hastie and Tibshirani, 1998; Fürnkranz, 2002), and error-correcting output codes (Dietterich and Bakiri, 1995). A general framework for such techniques is presented in (Allwein et al., 2000).

The main goal of these methods is to enable machine learning methods which are inherently designed for binary problems (e.g., perceptrons, support vector machines, etc.) to solve multi-class problems. However, there is also evidence that ensembles of binary classifiers may improve the performance of multi-class learners (Dietterich and Bakiri, 1995; Fürnkranz, 2003).

There are several reasons why such approaches can work. First, the binary problems are typically less complex and will often have a simpler decision boundary that is easier to model. For example, Knerr et al. (1992) observed that the classes of a digit recognition task were pairwise linearly separable, while it was not possible to discriminate each class from all other classes with linear perceptrons. It is well-known that Naive Bayes is essentially a linear classifier (Duda and Hart, 1972), and thus it can be expected to profit from such a pairwise decomposition of the task. Furthermore, a large number of binary classifiers introduces an ensemble effect in the sense that mistakes of a single classifier have a smaller impact on the final predictions, thus increasing the robustness of the classifier. For Naive Bayes classifiers in particular, which is known for giving good predictions but uncalibrated probabilities, class binarization is of importance because most calibration techniques operate on two-class problems (Zadrozny and Elkan, 2002). Finally, regarding computational efficiency, training a large number of classifiers from subsets of the training examples may be cheaper than training an entire classifier, in particular when the base classifier has a super-linear time or space complexity.

4 Pairwise Bayesian Classification

We are primarily interested in *pairwise classification*, which transforms an m-class problem into $m(m-1)/2$ two-class problems $\langle i, j\rangle$, one for each pair of classes $\{i, j\}$, $1 \leq i < j \leq m$. The binary classifier for problem $\langle i, j\rangle$ is trained with examples of classes c_i and c_j, whereas examples of classes $k \neq i, j$ are ignored for this problem. At classification time, a query x is submitted to all

binary models, and the predictions of the binary classifiers are combined to yield a final overall prediction.

A pairwise probabilistic classifier, R_{ij}, is therefore trained to estimate *pairwise probabilities* of the form

$$p_{ij} \;=\; \Pr(c_i|x, c_{ij}),$$

that is, the probability of class c_i given that the example x either belongs to class c_i or c_j (abbreviated as c_{ij}). These probabilities can, for example, be estimated by training a Bayes classifier on training sets D_{ij} which only contain the examples of classes c_i and c_j. More specifically, such a Bayes classifier estimates the pairwise probabilities $\Pr(c_i|x, c_{ij})$ and $\Pr(c_j|x, c_{ij})$ of class pair c_{ij} as follows:

$$\Pr(c_i|x, c_{ij}) = \frac{\Pr(x|c_i, c_{ij}) \cdot \Pr(c_i|c_{ij})}{\Pr(x|c_i, c_{ij}) \cdot \Pr(c_i|c_{ij}) + \Pr(x|c_j, c_{ij}) \cdot \Pr(c_i|c_{ij})}$$
$$\Pr(c_j|x, c_{ij}) = 1 - \Pr(c_i|x, c_{ij})$$

Again, a naive implementation of a Bayes classifier expands $\Pr(x|c_i, c_{ij})$ into $\Pr(a_1|c_i, c_{ij}) \cdot \Pr(a_2|c_i, c_{ij}) \cdots \Pr(a_m|c_i, c_{ij})$.

4.1 Bayesian Combination of Votes

The probabilities $p_{ij} = \Pr(c_i|x, c_{ij})$ need to be combined into probabilities (scores) $s_i = \Pr(c_i|x)$, a process that is known as *pairwise coupling* (Hastie and Tibshirani, 1998; Wu et al., 2004). In particular, we will consider simple *linear combiners* of the form

$$s_i \;=\; \sum_{j \neq i} w_{ij} \cdot \Pr(c_i|x, c_{ij}) \tag{3}$$

Interestingly, linear combination of that kind is sufficient to imitate regular Bayes classification:

Theorem 1. *Weighting the pairwise probabilities with*

$$w_{ij} = \frac{\Pr(c_{ij}|x)}{m-1} = \frac{\Pr(c_i|x) + \Pr(c_j|x)}{m-1} \tag{4}$$

reduces a pairwise Bayes classifier to a regular Bayes classifier.

Proof. Noting that

$$\begin{aligned}(m-1)\Pr(c_i|x) &= \sum_{j \neq i} \Pr(c_i|x) \\ &= \sum_{j \neq i} \Pr(c_i|x, c_{ij}) \cdot \Pr(c_{ij}|x),\end{aligned}$$

replacing w_{ij} in (3) by (4) yields

$$\begin{aligned} s_i &= \sum_{j \neq i} w_{ij} \cdot \Pr(c_i|x, c_{ij}) \\ &= \frac{1}{m-1} \sum_{j \neq i} \Pr(c_i|x, c_{ij}) \cdot \Pr(c_{ij}|x) \\ &= \frac{1}{m-1} \sum_{j \neq i} \Pr(c_i|x) \\ &= \Pr(c_i|x) = p_i \end{aligned}$$ □

This result is interesting, as it shows that an optimal Bayes decision can in principle be derived from an ensemble of pairwise learners. Or, stated differently, the binary decomposition of the original multi-class problem does not cause a loss of information. Moreover, the result shows how the weights w_{ij} should ideally be defined. As will be seen later on, the voting methods commonly used in practice refer to more simple weighting schemes, which can hence be considered as approximations of the ideal weights w_{ij}.

Anyway, as mentioned previously, the main interest in classification does not concern the probability estimates $\Pr(c_i|x, c_{ij})$ themselves, but only the resulting predictions. In the following, we will show that the use of voting or weighted voting will yield the same predictions as the Naive Bayes classifier.

4.2 Weighted Voting

Weighted Voting simply sums up the pairwise probability estimates of each class, i.e., it uses $w_{ij} \equiv 1$ in (3):

$$c_{WV} = \arg\max_{c_i} \sum_{j \neq i} \Pr(c_i|x, c_{ij})$$

Weighted voting has been frequently used in empirical studies and maintained a good performance. More importantly, Hüllermeier and Fürnkranz (2004) have shown that pairwise classification with weighted voting optimizes the Spearman rank correlation between the predicted ranking of all class labels and the true ranking, given that the predicted pairwise probabilities are unbiased estimates of their true values.

Theorem 2. *A pairwise Naive Bayes classifier with* weighted voting *predicts the same class ranking as a regular Naive Bayes classifier, i.e.,*

$$\Pr(c_i|x) \leq \Pr(c_j|x) \Leftrightarrow \sum_{k \neq i} \Pr(c_i|x, c_{ik}) \leq \sum_{k \neq j} \Pr(c_j|x, c_{jk})$$

$$\Pr(c_i|x) < \Pr(c_j|x) \Leftrightarrow \sum_{k \neq i} \Pr(c_i|x, c_{ik}) < \sum_{k \neq j} \Pr(c_j|x, c_{jk})$$

Proof. Let $p_i = \Pr(c_i|x)$ and $s_i = \sum_{k \neq i} \Pr(c_i|x, c_{ik})$. Then

$$\begin{aligned} s_i - s_j &= (p_{ij} - p_{ji}) + \sum_{k \neq i,j} p_{ik} - p_{jk} \\ &= \frac{p_i - p_j}{p_i + p_j} + \sum_{k \neq i,j} \frac{p_k(p_i - p_j)}{(p_i + p_k)(p_j + p_k)} \end{aligned}$$

From this, it immediately follows that $(p_i < p_j) \Rightarrow (s_i < s_j)$, and that $(p_i \leq p_j) \Rightarrow (s_i \leq s_j)$. The other directions can thus be obtained by contraposition: $(s_i < s_j) \Rightarrow (p_i < p_j)$ and $(s_i \leq s_j) \Rightarrow (p_i \leq p_j)$. □

Obviously, due to their construction, the pairwise probabilities p_{ij} are in full agreement with the total order induced by the regular Bayes probabilities p_i. In particular, it is interesting to note that these probabilities satisfy a certain type of *transitivity property*, namely

$$(p_{ij} < 1/2) \wedge (p_{jk} < 1/2) \;\Rightarrow\; (p_{ik} < 1/2). \tag{5}$$

This obviously holds, since $p_{ij} < 1/2$ means that $p_i < p_j$, $p_{jk} < 1/2$ means that $p_j < p_k$, and therefore $p_i < p_k$, which in turn implies $p_{ik} < 1/2$. It deserves mentioning, however, that, for many other pairwise base classifiers, this type of transitivity is not guaranteed. In fact, it is well possible that classifier R_{ik} predicts class c_k, classifier R_{kj} predicts class c_j, but classifier R_{ij}, predicts class c_i, resulting in a tie between the three classes. In fact, for rule learning algorithms not even symmetry will hold, i.e., $R_{ij} \neq R_{ji}$ (Fürnkranz, 2002).

On the other hand, one should also note that transitivity of a pairwise classifier is not enough to imply equivalence to the original (m-class) classifier. For example, suppose that $p_1 = 0.6$, $p_2 = 0.3$, $p_3 = 0.1$. The pairwise classifier with $p_{12} = p_{13} = 0.6$ and $p_{23} = 0.9$ is clearly transitive in the sense of (5) and also in agreement with the ranking $c_1 \succ c_2 \succ c_3$. Still, weighted voting gives $s_1 = 1.2$, $s_2 = 1.3$, $s_3 = 0.5$, and therefore the ranking $c_2 \succ c_1 \succ c_3$.

4.3 Unweighted Voting

An even simpler combination scheme for the predictions of pairwise classifiers is *unweighted voting*. To classify a new example, each of the learned base classifiers determines which of its two classes is the more likely one. The winning class receives a vote, and the algorithm eventually predicts the class that accumulates the highest number of votes. Essentially, it adds up the number of cases where class c_i has a higher pairwise probability than some other class c_j, i.e., the number of indexes j such that $\Pr(c_i|x, c_{ij}) \geq 0.5$ holds:

$$c_V = \arg\max_{c_i} \sum_{j \neq i} [\Pr(c_i|x, c_{ij})] = \sum_{j \neq i} \frac{[\Pr(c_i|x, c_{ij})]}{\Pr(c_i|x, c_{ij})} \cdot \Pr(c_i|x, c_{ij})$$

where $[x]$ is the rounding operator that returns $\lceil x \rceil$ if $x \geq \lfloor x \rfloor + 1/2$ and $\lfloor x \rfloor$ otherwise; thus, $w_{ij} = \frac{[\Pr(c_i|x, c_{ij})]}{\Pr(c_i|x, c_{ij})}$ in (3).

Again, we can show that this algorithm is equivalent to a regular Naive Bayes learner.

Theorem 3. *A pairwise Naive Bayes classifier with* unweighted voting *predicts the same class ranking as a regular Naive Bayes classifier.*

Proof. Let $p_i = \Pr(c_i|x)$ and $p_{ij} = \Pr(c_i|x, c_{ij}) = p_i/(p_i+p_j)$. Ignoring the issue of ties (i.e., $p_i \neq p_j$ for all $i \neq j$), one obviously has $(p_i < p_j) \Leftrightarrow (p_{ij} < p_{ji})$. Therefore, the number of votes received by class c_i is

$$s_i \;=\; \sum_{k \neq i} [p_{ik}]$$

and just corresponds to the number of classes c_k such that $p_i > p_k$. Therefore, the class with the k-th highest probability receives $m - k$ votes, which in turn means that the two rankings are identical. □

Remark 1. The above result can easily be generalized to the case of ties: If one splits up a vote in the case $p_{ij} = 1/2$, i.e., both classes receive one half of the vote, then $s_i = s_j$ iff $p_i = p_j$.

From the above proof it becomes immediately clear that the result in principle holds for all probabilistic or scoring classifiers that are "class-order-invariant" in the following sense: A class c_i receives a higher score than class c_j in the original m-class problem if and only if it is also favored in the direct (pairwise) comparison with c_j. In other words, the pairwise order between c_i and c_j is not reversed due to the consideration of additional classes, i.e., it is not influenced by the complementary set of classes $C \setminus \{c_i, c_j\}$. This property, which is quite interesting by itself, is obviously satisfied by a Bayes classifier but does not necessarily hold for other classifiers. Note that a class-order-invariant classifier also satisfies the transitivity property (5).

The above result is also interesting in light of the well-known robustness of Naive Bayes classification. As it shows in a rather explicit way, the main prerequisite for correct classification or, more generally, ranking of class labels, is not a very accurate estimation of probabilities. In fact, these probabilities are used by the pairwise classifier only in a very crude way, namely in the form of binary votes. Therefore, the only important thing is to correctly decide which among two given classes is the more probable one.

5 One-Against-All Class Binarization

In the previous section, we have seen that three versions of pairwise Naive Bayes classification are equivalent to a regular Naive Bayes classifier. At first sight, this is somewhat surprising, because what the Naive Bayes classifier does is modeling separate probability distributions for each individual class c_i first, and predicting the class with the maximum probability afterward. Thus, it seems to have much more in common with one-against-all classifiers than with pairwise

decomposition. However, it is not difficult to see that just the opposite is true, at least for the naive implementation of Bayes classification.

In one-against-all classification, an m-class problem is split into m binary problems that discriminate one class c_i, $i = 1 \dots m$, from all other classes. These classifiers are trained using all examples of class c_i as positive examples and the examples of the union of all other classes $\overline{c}_i = \bigcup_{j \neq i} c_j$ as negative examples. If we compare the two probabilities

$$\Pr(c_i|x)_{OA} = \frac{\Pr(x|c_i) \cdot \Pr(c_i)}{\Pr(x|c_i) \cdot \Pr(c_i) + \Pr(x|\bigcup_{j \neq i} c_j) \cdot \Pr(\bigcup_{j \neq i} c_j)}$$

$$\Pr(c_i|x)_{NB} = \frac{\Pr(x|c_i) \cdot \Pr(c_i)}{\Pr(x|c_i) \cdot \Pr(c_i) + \sum_{j \neq i} \Pr(x|c_j) \cdot \Pr(c_j)}$$

we can see that the difference lies in the normalizing constant, which in the case of the one-against-all Naive Bayes classifier estimates the probabilities from the sum of all counts over all classes $c_j, j \neq i$, whereas the regular Naive Bayes classifier sums the probabilities over these classes.

Since the equality relation

$$\Pr\left(x|\bigcup_{j \neq i} c_j\right) \cdot \Pr\left(\bigcup_{j \neq i} c_j\right) = \sum_{j \neq i} \Pr(x|c_j) \cdot \Pr(c_j)$$

generally holds, there is indeed no difference for a true Bayesian classifier. However, this equality is not valid for the probability estimates that are derived by Naive Bayes. If we use

$$f(c) = \Pr(c) \prod_{k=1}^{n} \Pr(a_k|c)$$

to denote the score that Naive Bayes computes for each class c, then

$$f(\bigcup_{j \neq i} c_j) \neq \sum_{j \neq i} f(c_j)$$

and, consequently, $\Pr(c_i|x)_{OA} \neq \Pr(c_i|x)_{NB}$. In particular, the probabilities $\Pr(c_i|x)_{OA}$ will in general not sum up to 1 ($\sum_i \Pr(c_i|x)_{OA} \neq 1$, but instead $\Pr(c_i|x)_{OA} + \Pr(\overline{c}_i|x)_{OA} = 1$ for all $i = 1, \dots, m$).

To see that this may also lead to different classifications (rankings of class labels), let us consider a sample problem with three classes A, B, and C, and 10 examples for each of them. We have two binary attributes X and Y. For $X = 1$ we have observed 15 examples distributed as $(1, 10, 4)$. Likewise, for $Y = 1$, we have 12 examples distributed as $(8, 1, 3)$. This gives

$$f(A) = \Pr(A) \cdot \Pr(X = 1|A) \cdot \Pr(Y = 1|A) = \frac{1}{3} \cdot \frac{1}{10} \cdot \frac{8}{10} = \frac{2}{75}$$

$$f(\overline{A}) = \Pr(\overline{A}) \cdot \Pr(X = 1|\overline{A}) \cdot \Pr(Y = 1|\overline{A}) = \frac{2}{3} \cdot \frac{14}{20} \cdot \frac{4}{20} = \frac{7}{75}$$

Analogously, we get

$$f(B) = \frac{1}{30}, \quad f(\overline{B}) = \frac{11}{120}; \qquad f(C) = \frac{1}{25}, \quad f(\overline{C}) = \frac{33}{200}$$

For a regular Naive Bayes, normalization yields

$$\Pr(A|X=1,Y=1)_{NB} = \frac{f(A)}{f(A)+f(B)+f(C)} = \frac{4}{15},$$
$$\Pr(B|X=1,Y=1)_{NB} = \frac{5}{15}; \quad \Pr(C|X=1,Y=1)_{NB} = \frac{6}{15},$$

and therefore the prediction C, whereas

$$\Pr(A|X=1,Y=1)_{OA} = \frac{f(A)}{f(A)+f(\overline{A})} = \frac{8}{36},$$
$$\Pr(B|X=1,Y=1)_{OA} = \frac{8}{30}; \quad \Pr(C|X=1,Y=1)_{OA} = \frac{8}{41},$$

and class B is predicted.

6 Discussion

The results obtained in this work, showing that various pairwise versions of a Naive Bayes classifier are equivalent to a regular Naive Bayes classifier, are interesting for several reasons. As a first consequence, decomposing a multi-class problem into a pairwise ensemble of binary classifiers does not work for Naive Bayes classifiers, that is, it is not possible no improve classification performance in this way.

The weights derived in Theorem 1 are not specific to Naive Bayes, but apply to probabilistic algorithms in general. It remains to be seen whether this technique can be applied to other base classifiers as well. The main practical impediment is the estimation of $\Pr(c_{ij}|x)$. For example, one could try to estimate them using a pairwise variant of Naive Bayes that predicts a *pair* of classes instead of a single class. First experiments with a few related variants, presented in (Sulzmann, 2006), were not very encouraging, however.

Hüllermeier and Fürnkranz (2004) have shown that weighted voting optimizes the Spearman rank correlation, provided the pairwise probabilities are estimated correctly. In this work, we have shown the equivalence of Naive Bayes to pairwise Naive Bayes using weighted voting. Combining these two results lets us conclude that the regular Naive Bayes also optimizes the Spearman rank correlation. The main problem, of course, is that its probability estimation is biased because of the independence assumption, which will in general not hold. However, just as the bias in the probability estimation does not necessarily affect the prediction of the top rank (Domingos and Pazzani, 1997), it might well turn out that its effect on the entire ranking of the classes is not very strong; the equivalence result for pairwise Naive Bayes with unweighted voting in Section 4.3 is clearly

indicative in this regard, as is the generally good performance of Naive Bayes on ranking tasks (Zhang and Su, 2004). We plan to elaborate on this issue in future work.

Another interesting issue concerns the generalization of the results obtained in this paper. For example, we already mentioned that the equivalence between regular Bayes and pairwise Bayes with unweighted voting in principle holds for all "class-order-invariant" classifiers. Finding a characterizing property of a similar kind appears to be more difficult in the case of weighted voting. For Bayes classification, there is a very simple relationship between the multi-class probabilities p_i and the pairwise probabilities p_{ij}: the latter are directly proportional to the former. As we have seen, this relationship assures that the order of the classes remains unchanged, that is, this property is sufficient to guarantee equivalence. However, it is presumably not a necessary condition.

Acknowledgments

This research was supported by the *German Science Foundation (DFG)*.

References

Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research 1, 113–141 (2000)

Anand, R., Mehrotra, K.G., Mohan, C.K., Ranka, S.: Efficient classification for multiclass problems using modular networks. IEEE Transactions on Neural Networks 6, 117–124 (1995)

Clark, P., Boswell, R.: Rule induction with CN2: Some recent improvements. In: Kodratoff, Y. (ed.) EWSL 1991. LNCS, vol. 482, pp. 151–163. Springer, Heidelberg (1991)

Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)

Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263–286 (1995)

Domingos, P., Pazzani, M.J.: On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning 29(2-3), 103–130 (1997)

Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley, New York (1972)

Friedman, J.H.: Another approach to polychotomous classification. Technical report, Department of Statistics, Stanford University, Stanford, CA (1996)

Fürnkranz, J.: Round robin classification. Journal of Machine Learning Research (JMLR) 2, 721–747 (2002)

Fürnkranz, J.: Round robin ensembles. Intelligent Data Analysis 7(5), 385–403 (2003)

Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems 10 (NIPS-97), pp. 507–513. MIT Press, Cambridge (1998)

Hüllermeier, E., Fürnkranz, J.: Ranking by pairwise comparison: A note on risk minimization. In: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE-04), Budapest, Hungary (2004)

Kononenko, I.: Inductive and bayesian learning in medical diagonsis. Applied Artificial Intelligence 7(4), 331–337 (1993)

Knerr, S., Personnaz, L., Dreyfus, G.: Handwritten digit recognition by neural networks with single-layer training. IEEE Transactions on Neural Networks 3(6), 962–968 (1992)

Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research 5, 101–141 (2004)

Sulzmann, J.-N.: Pairwise Naive Bayes classifier. In: Althoff, K.-D., Schaaf, M. (eds.) Proceedings of the LWA 2006, Lernen Wissensentdeckung Adaptivität, Hildesheim, Germany, pp. 356–363. Gesellschaft für Informatik e. V (GI) (2006)

Wu, T.-F., Lin, C.-J., Weng, R.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research (JMLR) 5, 975–1005 (2004)

Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD-02), pp. 694–699 (2002)

Zhang, H., Su, J.: Naive Bayesian Classifiers for Ranking. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 501–512. Springer, Heidelberg (2004)