On the Combination of Locally Optimal Pairwise Classifiers

G. Szepannek *, B. Bischl and C. Weihs

Department of Statistics, Dortmund University of Technology, D-44221 Dortmund

Abstract

Classification methods generally rely on some idea about the data structure. If the specific assumptions are not met, a classifier may fail. In this paper the possibility of combining classifiers in multi-class problems is investigated. Multi-class classification problems are split into two class problems. For each of the latter problems an optimal classifier is determined. The results of applying the optimal classifiers on the two class problems can be combined using a *pairwise coupling* algorithm.
In this paper exemplary situations are investigated where the respective assumptions of Naive Bayes or the classical Linear Discriminant Analysis (LDA) fail. It is investigated at which degree of violations of the assumptions it may be advantageous to use single methods or a classifier combination by pairwise coupling.

Key words: Multi-class classification, combining classifiers, pairwise coupling

1 Introduction

Classification methods generally rely on some idea about the data structure, i.e. some assumptions are made to be valid for the distribution of the classes. If these assumptions are not met, classifiers may fail. Based on work presented at the MLDM 2007 conference [Szepannek et al., 2007], the possibility of combining classifiers in multi-class problems is investigated in this paper. When talking about ensemble methods one usually has in mind popular principles like *bagging* [Breiman, 1996] or *boosting* [Freund and Shapire, 1997]. A well known example of bagging is to improve the performance of single decision trees by *random forests* [Breiman, 2001]. [Prinzie and van der Poel, 2006] use a similar approach for multinomial logistic regression. Both bagging and

* Corresponding author.
Email address: szepannek@statistik.tu-dortmund.de (G. Szepannek).

boosting rely on combinations of different classification rules that are built on sampled or weighted instances of the original data. In [Wozniak, 2006] a method is proposed to combine several classifiers by weighted voting.
In this paper a somewhat different perspective to combining classifiers for multi-class problems is worked out: the basic observation is that classifiers only work well if their underlying assumptions hold, e.g. class wise independent features in the case of a *Naive Bayes* classifier. This might be true for some but not necessarily all of the classes.
Multi-class classification problems are split into two class problems. For each of the latter problems an optimal classifier is determined. The results of applying the optimal classifiers on the two class problems can be combined using a *pairwise coupling* algorithm [Hastie and Tibshirani, 1998] which is presented in Section 3.
In this paper, the principle of combining pairwise optimal classifiers is investigated for the case of two very common classification methods, namely *Naive Bayes* and *Linear Discriminant Analysis* [Fisher, 1936]. Both methods are briefly described in Section 4. In a simulation study in Section 5 the degree of violation of the assumption of both methods is varied. The results give quite an interesting indication of the robustness of both methods as well as they produce a 'map' that shows when to use whether one of the single classifiers or a combination. It turns out that in some situations a combination of pairwise optimized classifiers can strongly improve the classification results if the assumptions of single methods do not hold for all classes. Section 6 shows that this results has also been observed for application to several real world classification problems.

2 Classification Algorithm

Pairwise coupling (PWC) generates $K(K-1)/2$ subsamples of the data (K being the number of classes) each consisting only of objects of one specific *pair of classes*. For these two classes, an optimal classifier is determined, e.g. using cross-validation. According to the thoughts presented above, the optimal classifier may be a different one for different class pairs. A similar approach is made in [Szepannek and Weihs, 2006] to perform classification on locally optimally selected feature subspaces.
Prediction models for all class pairs can be applied but in general, when classifying new data no prior information is available to which pair of classes an object belongs. To solve this problem, the *pairwise coupling* algorithm [Hastie and Tibshirani, 1998] can be employed for construction of class membership posterior probabilities (and thus a multi-class classification rule) from the (pairwise) results.

The following pseudo-code summarizes the steps of the suggested proceeding:

Build classification model *(data, set of classification methods)*

(1) For each pair of two classes do
(2) (a) Remove temporarily all observations that do not belong to one of both classes from *data*: return *newdata*.
    (b) For each *classifier* in *set of classification methods*
        - Build *classifier* on *newdata*.
        - Validate *classifier* e.g. using cross-validation.
        - Store Results temporally in *classifier results*.
    (c) Choose best *classifier* according to *classifier results* return *classifier of class-pair*.
    (d) Train *classifier of class-pair* on *newdata*:
    (e) Return *model of class-pair*.
(3) Return the whole model consisting of *model of class-pair* for all pairs of classes.

Predict class *(new object, models of class-pairs)*

(1) For each pair of subclasses do
(2) (a) Calculate the posterior probabilities for *new object* assuming the object being of one of the currently considered two classes according to *model of class-pair*.
    (b) Return the *class pair posterior probabilities*.
(3) Use the pairwise coupling algorithm to calculate the posterior probabilities for all K classes from the set of all estimated pairs of conditional *class pair posterior probabilities*.
(4) Return the predicted class k with maximal *class posterior probability*.

The next section describes a solution to the problem of gaining the vector of posterior probabilities form the pairwise classification models built with possibly different classifiers.

3 Pairwise Coupling

3.1 Definitions

We now tackle the problem of finding posterior probabilities of a K-class classification problem given the posterior probabilities for all $K(K-1)/2$ pairwise comparisons. Let us start with some definitions.

Let $p(x) = p = (p_1, \ldots, p_K)$ be the vector of (unknown) posterior probabilities. p depends on the specific realization x. For simplicity in notation we will omit x. Assume the "true" conditional probabilities of a pairwise classification problem to be given by

$$\mu_{ij} = Pr(i|i \cup j) = \frac{p_i}{p_i + p_j} \quad . \tag{1}$$

Let r_{ij} denote the estimated posterior probabilities of the two-class problems. The aim is now to find the vector of probabilities p_i for a given set of values r_{ij}.

Example 1:

Let $p = (0.7, 0.2, 0.1)$. The μ_{ij} can be calculated according to equation 1 and can be presented in a matrix:

$$(\mu_{ij})_{i,j} = \begin{pmatrix} . & 7/9 & 7/8 \\ 2/9 & . & 2/3 \\ 1/8 & 1/3 & . \end{pmatrix} \quad . \tag{2}$$

Example 2:

The inverse problem does not necessarily have a proper solution, since there are only $K - 1$ free parameters but $K(K-1)/2$ constraints. Consider

$$(r_{ij})_{i,j} = \begin{pmatrix} . & 0.9 & 0.4 \\ 0.1 & . & 0.7 \\ 0.6 & 0.3 & . \end{pmatrix} \tag{3}$$

where the row i contains the estimated conditional pairwise posterior probabilities r_{ij} for class i. It can be easily checked that the linear system resulting from applying equation 1 cannot be solved.
From Machine Learning, majority voting ("Which class wins most comparisons ?") is a well known approach to solve such problems. But here, it will not lead to a result since any class wins exactly one comparison. Intuitively, class 1 may be preferable since it dominates the comparisons the most clearly.

3.2 Algorithm

In this section we present the pairwise coupling [Hastie and Tibshirani, 1998] to find p for a given set of r_{ij}. They transform the problem into an iterative optimization problem by introducing a criterion to measure the fit between the observed r_{ij} and the $\hat{\mu}_{ij}$, calculated from a possible solution $\hat{p}$. To measure the fit they define the weighted Kullback-Leibler distance:

$$l(\hat{p}) = \sum_{i<j} n_{ij} \left(r_{ij} \ln \left(\frac{r_{ij}}{\hat{\mu}_{ij}} \right) + (1 - r_{ij}) \ln \left(\frac{1 - r_{ij}}{1 - \hat{\mu}_{ij}} \right) \right) \quad . \tag{4}$$

n_{ij} is the number of objects that fall into one of the classes i or j.
The best solution $\hat{p}$ of posterior probabilities is found as in iterative proportional scaling (IPS) (for more details on the optimization see [Bishop et al., 1975] or [Herbert, 1988]). The algorithm consists of the following three steps:

(1) Start with any $\hat{p}$ and calculate all $\hat{\mu}_{ij}$.
(2) Repeat until convergence $i = (1, 2, \ldots, K, 1, \ldots)$:

$$\hat{p}_i \leftarrow \hat{p}_i \frac{\sum_{j \neq i} n_{ij} r_{ij}}{\sum_{j \neq i} n_{ij} \hat{\mu}_{ij}} \quad , \tag{5}$$

renormalize $\hat{p}$ and calculate the new $\hat{\mu}_{ij}$.
(3) Finally scale the solution to $\hat{p} \leftarrow \frac{\hat{p}}{\sum_i \hat{p}_i}$.

Motivation of the algorithm: It has been shown [Hastie and Tibshirani, 1998] that $l(p)$ increases at each step. For this reason, since it is bounded above by 0, $l(p)$ converges. providing $\hat{\mu}_{ij} = r_{ij} \; \forall \, i \neq j$, it will be found. Even if the choice of $l(p)$ as optimization criterion is rather heuristic, it can be motivated in the following way: consider a random variable $n_{ij} r_{ij}$, being the number of observations of class i among the n_{ij} observations of class i and j. This random variable can be considered to be binomially distributed $n_{ij} r_{ij} \sim B(n_{ij}, \mu_{ij})$ with "true" (unknown) parameter μ_{ij}. Since the same (training) data is used for all pairwise estimates r_{ij}, the r_{ij} are not independent, but if they were, $l(p)$ of equation 4 would be equivalent to the log-likelihood of this model (see [Bradley and Terry, 1952]). Then, maximizing $l(p)$ would correspond to maximum-likelihood estimation for μ_{ij}.
Going back to example 2, we obtain $\hat{p} = (0.47, 0.25, 0.28)$, a result being consistent with the intuition that class 1 may be slightly preferable.
In [Wu et al., 2004] several methods for multi-class probability estimation by pairwise coupling algorithms are presented and compared. For the simulations of this paper, the method of [Hastie and Tibshirani, 1998] is used.

4 Implemented Methods

4.1 Linear Discriminant Analysis

In its classical form *Linear Discriminant Analysis* (LDA) [Fisher, 1936] was constructed for linear reduction of dimensionality to maximize the distance of class means w.r.t. the covariance structure of the data.
The method is shown to be optimal in the sense that it minimizes the *Bayes*

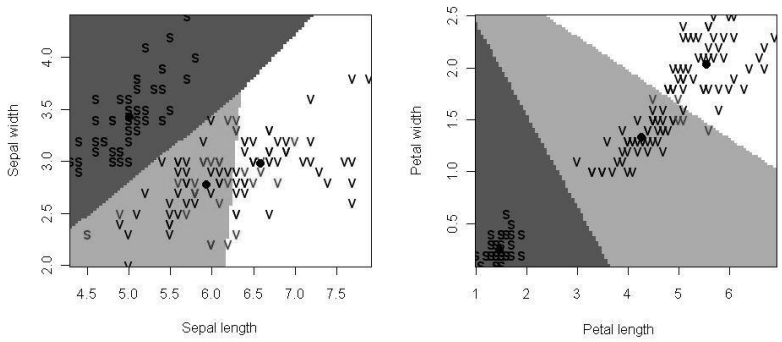

Fig. 1. Two-dimensional projections of the partition of the feature space using Linear Discriminant Analysis on Iris data.

risk if the underlying class distributions follow normal law but have equal covariance matrices for all classes (see e.g. [Hastie et al., 2001], p.95).
The classification for an object x is obtained by maximizing the decision rule $\hat{d}_k(x)$ over all classes k:

$$\hat{d}_k(x) = \bar{x}_k \hat{\Sigma}^{-1} x - \frac{1}{2} \bar{x}_k \hat{\Sigma}^{-1} \bar{x}_k + \ln(\pi(k)) \tag{6}$$

with $\pi(k)$ being the class prior membership probabilities, $\bar{x}_k$ denoting the mean of class k and $\hat{\Sigma}$ being the *pooled covariance matrix*

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{n=1}^{N} I_{[k]}(k_n)(x_n - \bar{x}_k)(x_n - \bar{x}_k)'. \tag{7}$$

Here $I_{[k]}(k_n)$ represents the indicator function that becomes 1 if object n of the training data is of class k and 0 if not. The term pooled covariance follows from the fact that equation (7) can be reformulated in terms of the classwise

covariance estimations $\hat{\Sigma}_k$:

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} n_k \hat{\Sigma}_k \tag{8}$$

where n_k denotes the size of class k in the training data.
The classification rule linearly partitions the feature space. This is shown in Figure 1 for the first two dimensions of the well known iris data from [Fisher, 1936]. In [Hastie et al., 2001] it is mentioned that Linear Discriminant Analysis often shows good results and is among top 3 classifiers for 7 of 22 real world data data sets of the Statlog project [Michie et al., 1994].

4.2 Naive Bayes

When using the *Naive Bayes* method features are assumed to be conditionally independent given the class. For each class k and variable d mean $\hat{\mu}_{d,k}$ and covariance $\hat{\sigma}_{d,k}$ are estimated.
For a new observation x the likelihood $P_d(x|k)$ of its realization in variable d

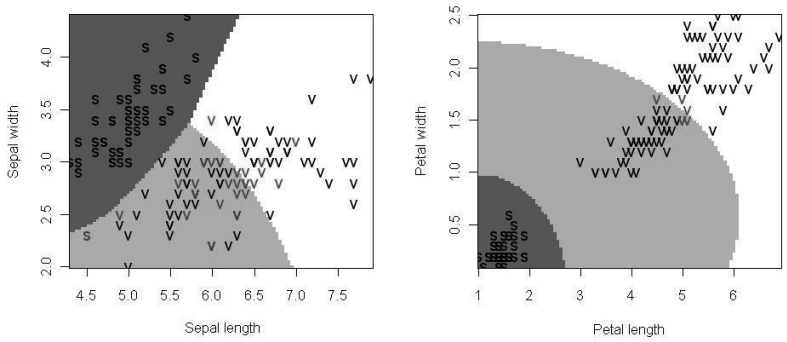

Fig. 2. Two-dimensional projections of the partition of the feature space using Naive Bayes on Iris data.

given class k can be calculated then assuming normal distribution.
Finally, the predicted class is obtained my maximizing the decision rule

$$\hat{d}_k(x) = \pi(k) \prod_d P_d(x|k). \tag{9}$$

with $\pi(k)$ again denoting the prior probability of class k. Doing so implicitly assumes no correlations between the different variables d given the class: the covariance matrix of class k Σ_k is assumed to be 0 for all elements except for the main diagonal elements.

This dramatically decreases the number of free model parameters, especially if the number of features is large. Another advantage of the Naive Bayes method may be that equal variances are not assumed as it is done in LDA. Nevertheless, it may be disadvantageous if there are strong correlations among the predictor variables.

5 Simulation Study

5.1 An Introductory Example

To gain some insight into the merit of the method a synthetic example was constructed. This example consisted of four equally large classes in two-dimensional space, all normally distributed (see Fig. 3, the different classes are labelled with numbers from 1 to 4).

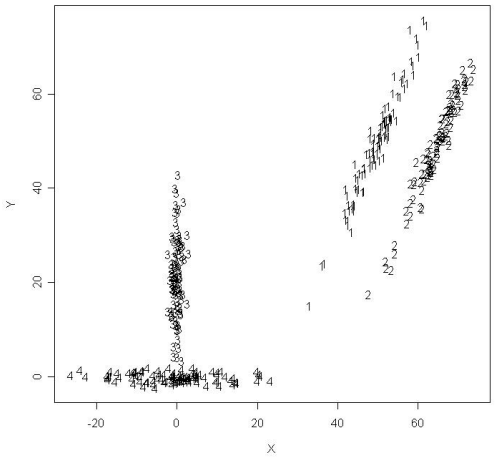

Fig. 3. First example of simulated data.

Classes 1 and 2 have an equal covariance structure and can therefore be optimally separated by an LDA classifier, but not by the Naive Bayes method since the input variables are not independent given the class.
Likewise, as classes 3 and 4 have uncorrelated features given the class, they can therefore be optimally identified by the Naive Bayes method, but LDA will produce a higher error because the underlying normal distributions do not have an equal covariance matrix.

It is now conjectured that by training a PWC classifier on the dataset, a LDA-classifier is chosen to separate the first pair of classes and a Naive Bayes classifier for the latter pair. This expected behaviour can be observed on the simulated data. The results show a strong increase in classification performance on separately simulated test data when combing both classifiers as

Method	Test Error
LDA	0.07
Naive Bayes	0.14
PWC	**0.01**

Table 1
Test error rates on synthetic example of Fig. 3 (400 samples per class, 2/3 training data and 1/3 test data).

opposed to use only the base methods (see table 3).

5.2 Experimental Setting

In oder to investigate when it is beneficial to use one of the base methods or their classification using pairwise coupling (and choosing the pairwise optimal classifier based on cross-validated error rates) a study is performed with simulated data as in Section 5.1 but with varying degree of violated assumptions for both methods:
Four normally distributed classes are generated with class expectations:

$$\mu_1 = (50, 50)', \ \mu_2 = (65, 50)', \ \mu_3 = (0, 20)', \ \mu_4 = (0, 0)'.$$

The class covariance matrices are constructed as a convex combination of four extreme cases:

$$\Sigma_1^*(\rho) = \Sigma_2^*(\rho) = \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\,\sigma_2 \\ \rho\,\sigma_1\,\sigma_2 & \sigma_2^2 \end{pmatrix},$$

$$\Sigma_3^*(\rho) = \Sigma_3^* = \begin{pmatrix} \sigma_3^2 & 0 \\ 0 & \sigma_4^2 \end{pmatrix} \quad \textit{and} \quad \Sigma_4^*(\rho) = \Sigma_4^* = \begin{pmatrix} \sigma_4^2 & 0 \\ 0 & \sigma_3^2 \end{pmatrix}$$

with $\sigma_1 = 5$, $\sigma_2 = 10$, $\sigma_3 = 1$ and $\sigma_4 = 10$.
The covariance matrices of class 1 and 2 exactly hold the assumptions that underly Linear Discriminant Analysis, since covariances are greater 0 but the same for both classes. The covariance matrices of class 3 and 4 represent the 'naive Bayes - case' since the variables are independent but have different

variances for both classes.
The covariance Σ_i of class i is set to be

$$\Sigma_i(\alpha, \rho) = \alpha\Sigma_1^*(\rho) + (1-\alpha)\Sigma_i^*(\rho) \tag{10}$$

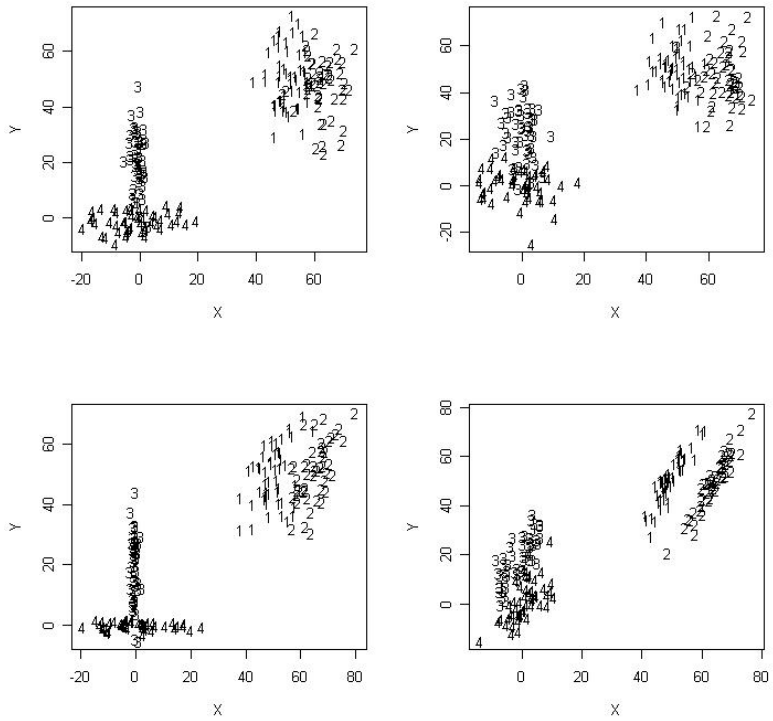

Fig. 4. Simulated data for 4 different parameter combinations.

The parameter $\alpha \in [0, 1]$ determines how equal the class covariance matrices look like, the larger α is the more equal they are. For $\alpha = 1$ all classes' covariances equal to $\Sigma_1^*(\rho)$. Then, the assumptions of LDA holds.
The free parameter $\rho \in [0, 1]$ determines the correlation in $\Sigma_1^*(\rho)$. $\rho = 0$ means independent variables for all classes as it is assumed for the Naive Bayes method. Four exemplary situations are shown in Figure 4: The upper left figure shows simulated data for $\alpha = 0.1, \rho = 0.1$: all classes possess quite specific covariance matrices with very small correlations among the variables. This should be a case where the Naive Bayes method can be assumed to produces good results. The upper right figure illustrates simulated data for $\alpha = 0.5, \rho = 0$: for all classes the variables are completely uncorrelated but the class-specific covariance structure is not as present as in the example before. The bottom left figure illustrates the data situation for $\alpha = 0, \rho = 0.5$: The covariance matrices of the classes are unique and the variables of class 3 and 4 are correlated. Both parameters are set to $\alpha = \rho = 0.9$ in the bottom right figure: The covariance shapes of the classes look very similar and contain strong correlations. In this situation, the assumptions of LDA are quite well

met.
For our simulation study both parameters α and ρ are varied in the interval $[0, 1]$. For each simulation 400 observations are generated for each class. The data are split into 2/3 training data. The last third is used for testing. The locally optimal classifiers are chosen by 3-fold cross-validation.

5.3 Results

The results of the simulation study are shown in Figure 5. The first three plots show the results of the two base methods as well as their combination using PWC for our simulated data. For each pair of simulation parameters (α, ρ)

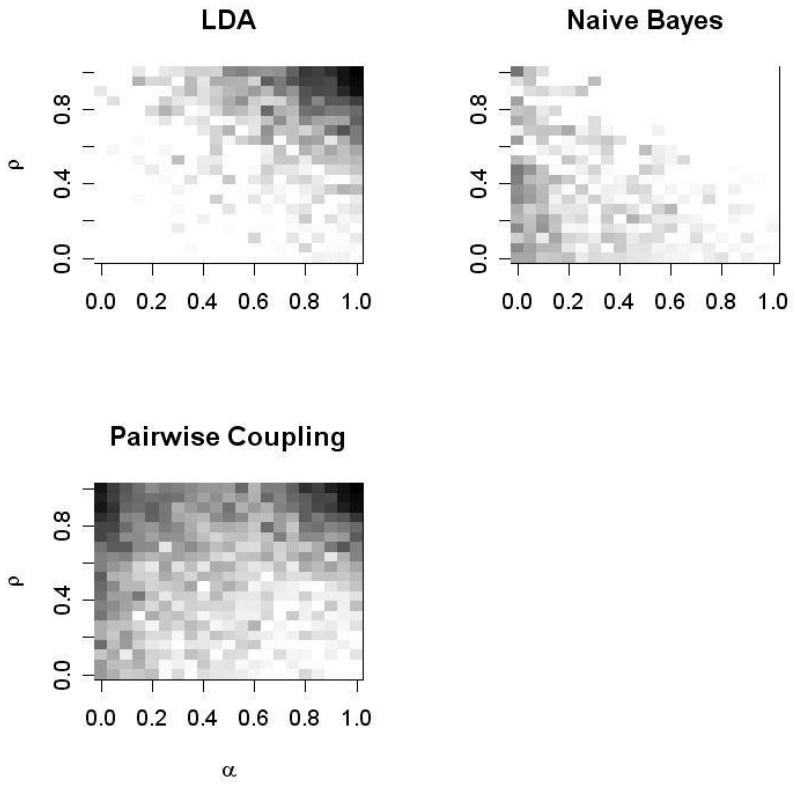

Fig. 5. Results of the simulation study for the different methods (scaled between 0 error (black) and the worst result (white).

the results are scaled: black indicates a test error rate of 0 while white denotes the worst obtained result.
It can be easily recognized that LDA performs best for both high parameters of ρ and α, i.e. equal covariance matrices of all classes and strong correlations between the variables. Using Naive Bayes is advantageous for a low parameter α, i.e. strongly differing covariance matrices of the classes, especially if there are furthermore low correlations in the variables. Combining both classifiers is

a good compromise in most situations except if there are equal covariance matrices with small correlations of all classes. A strong benefit can be obtained if the covariance matrices of the classes are not equal and there are also strongly correlated variables, i.e. if the assumptions of both base methods do not hold.
To determine whether one method significantly outperforms the other two,

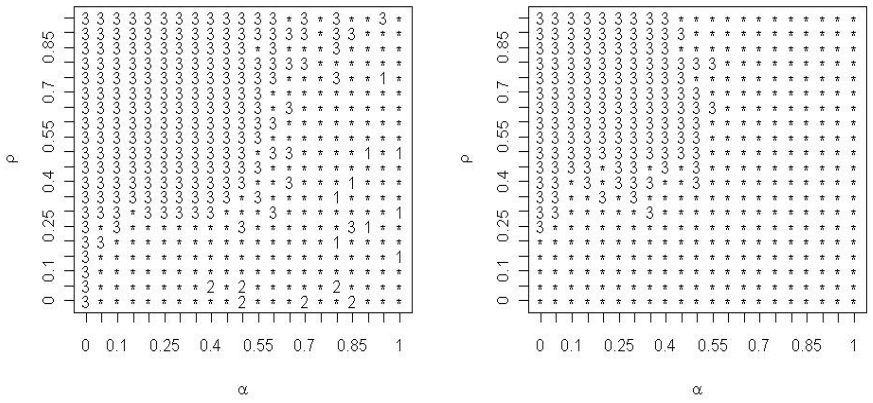

Fig. 6. Significantly best method in dependence of the parameter compared to the best competitor. * indicates that none of the methods significantly outperforms the other methods. Method 1: LDA; 2: Naive Bayes and 3: PWC. Left: simple significances, right: using Bonferroni-Holm correction for multiple testing.

the above mentioned simulation was conducted 30 times and for each pair of parameters a paired t-test between the test error rates winning and the second best method was applied [Dietterich, 1998]. Figure 6 shows the results. In the left figure tests were performed with a simple significance level of 0.05, while in the right figure - in order to cope with the problem of multiple testing - results are given after adapting the significance levels by the Bonferroni-Holm method. One can observe that there are situations (strongly differering covariances between the classes combined with correlations that appear in the data) where the PWC approach leads to significant improvements of the misclassification rate. The base methods show a possible advantage in the areas where their underlying assumptions are met (low ρ for Naive Bayes and and high α for LDA implying equal covariance matrices in the classes). But these advantages proved not to be significant for adapted significance levels and might be due to multiple testing.

6 Application to Real World Data

For not to restrict our analysis on the simulated data we also applied the methods to several real world multi-class data from the *UCI Machine Learn-*

ing Repository. An overview over some characteristics of the chosen data sets is given in Table 2. For an explicit description of the data sets see [Michie et al., 1994] and [Merz and Murphy, 1998].

In each experiment the data were randomly split into a training and test set

	Satellite	Vehicle	Nursery	Vowel
classes	6	4	5	11
features	36	18	8	10
examples	6435	846	12960	990

Table 2
Statistics of data sets.

(2/3 and 1/3), except for the Satellite set, where the same 4435 examples as in Statlog were used for training and the remaining 2000 examples for testing. The results are given in Table 3 in terms of test error rates for both base methods as well as their combination. For the Satellite data, the error rates of the Naive Bayes method can be improved by a combined classifier but LDA performs overall best. For the Vehicle data set, Naive Bayes shows very bad results. The rates of LDA can even be slightly improved using a PWC classifier combination. For the Nursery data LDA shows very bad results. The error rates of Naive Bayes here can be improved by pairwise coupling. Finally, for the Vowel data set the recognition rates of both methods can be dramatically improved using a classifier combination. As a conclusion, the proposed local combination of classifiers sometimes yielded a large improvement of the results but never showed very bad performance compared to the winning base method method. This result is in harmony with the observations made in Section 5.

Method	Satellite	Vehicle	Nursery	Vowel
LDA	**0.15**	0.26	0.47	0.42
Naive Bayes	0.20	0.57	0.10	0.52
PWC	0.18	**0.23**	**0.08**	**0.17**

Table 3
Test error rates on UCI real world data sets.

7 Summary

Classifier combination for multi-class classification problems is proposed in a different way compared to the very common bagging and boosting approaches: for each pair of classes an optimal classifier is determined using cross-validation and class pairwise models are trained. A new object is labelled by applying

all classifiers for each class pair and then combining the results by pairwise coupling [Hastie and Tibshirani, 1998].
Such a proceeding may be advantageous in situations where the assumptions of the different base methods hold for different classes.
The benefit of such a classifier combination is investigated for two very common methods, namely Linear Discriminant Analysis and Naive Bayes. A simulation study is performed where the degree of violation of the specific assumptions for both methods is varied and a map is obtained indicating when it is better to implement a single one of these methods or their combination. Furthermore, the methods are applied to common real world problems from the UCI Machine Learning Repository. It turned out that sometimes large improvements of the misclassification rate are achieved by using PWC while its results were never much worse than the winning base method.
It should also be mentioned that in [Moreira and Mayoraz, 1998] a different approach is proposed to build classifiers from class pairwise rules by calculating conditional probabilities for the membership of a new object to a class pair. A comparison to this approach may be a topic of further investigation as well as the investigation of the principle using other classifiers.
Finally – referring to the work of [Dietterich and Baikiri, 1995] – multiclass-classification problems can also be solved by transforming them into several binary classification problems using *Error-Correcting Output Codes*. There basically, in every binary classification problem the K classes are grouped into two sets of classes which are then separated. The result is a sequence of binary classifiers. Each of the classes is coded by a vector of the binary group-labels. Prediction of an object is done by applying all classifiers and choosing the class with the most similar code vector. In [Hülsmann and Friedrich, 2006] an application of several ECOC strategies to SVMs is shown.
PWC can be embedded in this context according to [Allwein et al., 2000] and thus an extension of the suggested approach towards Error-Correcting Output Codes may also be topic of further investigation.
Acknowledgment. This work has been supported by the Collaborative Research Center (SFB) 475 of the German Research Foundation (DFG).

References

[Allwein et al., 2000] Allwein, E., Schapire, R. and Singer, Y.:, 2000, Reducing multiclass to binary: a unifying approach for margin classifiers. Journal of Machine Learning Research **1**, 113–141.

[Bishop et al., 1975] Bishop, Y., Fienberg, S. and Holland, P., 1975, Discrete Multivariate Analysis. MIT Press, Cambridge.

[Bradley and Terry, 1952] Bradley, R. and Terry, M., 1952, The rank analysis of

incomplete block designs, i. the method of paired comparisons. Biometrics, 324–345.

[Breiman, 1996] Breiman, L., 1996, Bagging predictors. Machine Learning **24(2)**, 123–140.

[Breiman, 2001] Breiman, L., 2001, Random forests. Machine Learning **45(1)**, 5–32.

[Dietterich and Baikiri, 1995] Dietterich, T. and Bakiri, G., 1995, Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research **2**, 263–286.

[Dietterich, 1998] Dietterich, T., 1998, Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation **10 (7)**, 1895–1923.

[Fisher, 1936] Fisher, R., 1936, The use of multiple measures in taxonomic problems. Annals of Eugenics **7** , 179–188.

[Freund and Shapire, 1997] Freund, Y. and Schapire, R., 1997, A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55(1)**, 119–139.

[Hastie and Tibshirani, 1998] Hastie, T. and Tibshirani, R., 1998, Classification by pairwise coupling. Annals of Statistics **26(1)**, 451–471.

[Hastie et al., 2001] Hastie, T., Tibshirani, R. and Friedman, J. 2001, The Elements of Statistical Learning - Data Mining, Inference and Prediction. Springer, NY.

[Herbert, 1988] Herbert, D., 1988, *The Method of Paired Comparisons*. 2^{nd} edition, Charles Griffin, London.

[Hülsmann and Friedrich, 2006] Hülsmann, M. and Friedrich, M., 2007, Comparison of a novel combined ECOC strategy with different mutliclass algorithms together with parameter optimization methods. in: P.Perner (Ed.): *Machine Learning and Data Mining in Pattern Recognition*, LNAI 4571, Springer Verlag, Heidelberg, 17–31.

[Merz and Murphy, 1998] Merz, C. and Murphy, P., 1998, UCI repository of machine learning data bases: `http://www.ics.uci.edu/~mlearn/mlrepository.html`, Irvine, CA: University of California, Dept. of Information and Computer Science.

[Michie et al., 1994] Michie, D., Spiegelhalter, D. and Taylor, C., 1994, Machine Learning, Neural and Statistical Classification. Ellis Horwood Limited, Hertfordshire.

[Moreira and Mayoraz, 1998] Moreira, M. and Mayoraz, E., 1998, Improved pairwise coupling classification with correcting classifiers. European Conference on Machine Learning, 160–171.

[Prinzie and van der Poel, 2006] Prinzie, A. and van der Poel, D., 2006, Exploiting randomness for feature selection in multinomial logit: a CRM cross-sell application. in: P.Perner (Ed.): *Advances in Data Mining*, LNAI 4065, Springer Verlag, Heidelberg, 310–323.

[Szepannek et al., 2007] Szepannek, G., Bischl, B. and Weihs, C., 2007, On the combination of locally optimal pairwise classifiers. in: P.Perner (Ed.): *Machine Learning and Data Mining in Pattern Recognition*, LNAI 4571, Springer Verlag, Heidelberg, 104–116.

[Szepannek and Weihs, 2006] Szepannek, G. and Weihs, C., 2006, Local modelling in classification on different feature subspaces. in: P.Perner (Ed.): *Advances in Data Mining*, LNAI 4065, Springer Verlag, Heidelberg, 226–238.

[Wozniak, 2006] Wozniak, M., 2006, Adaptive weights calculation procedure for weighted voting – idea and experimental results. in: P.Perner (Ed.): Poster Proceedings of ICDM Conference, Leipzig, 2006.

[Wu et al., 2004] Wu, T., Lin, C. and Weng, R., 2004, Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research **5**, 975–1005.