Local Modelling in Classification on Different Feature Subspaces

Gero Szepannek Claus Weihs

Februar 2006

Fachbereich Statistik
Universität Dortmund

Abstract

Sometimes one may be confronted with classification problems where classes are constituted of several subclasses that possess different distributions and therefore destroy accurate models of the entire classes as one similar group. An issue is modelling via local models of several subclasses.
In this paper, a method is presented of how to handle such classification problems where the subclasses are furthermore characterized by different subsets of the variables. Situations are outlined and tested where such local models in different variable subspaces dramatically improve the classification error.

1 Introduction

In order to minimize the misclassification error in a $C-$class classification problem one aims at searching for a classification rule

$$\hat{c} = \arg \max_{c=1,\ldots,C} P(c|x) \tag{1}$$

that maximizes the conditional posterior probability given the observation x. It may be the case that a class c is composed of several ”subclasses” with different distributions. For an accurate estimation of $P(c|x)$ these subclasses have to

be modelled separately by *local models*. During this paper, we assume all the subclass-memberships in the training data to be known, whereas these memberships in the test data - of course - are not known (else the class of the observation would also be given!). If the subclasses are not known in advance clustering methods can be used to investigate if the data of some class is composed from several subgroups of data.
We call $k = \{1, \ldots, K\}$ the index of all subclasses. There is existing a (surjective) relationship $f : \{1, \ldots, K\} \rightarrow \{1, \ldots, C\}$. Given the posterior probabilities of the membership of any of the subclasses $P(k|x)$, the classification rule for any class c is given by

$$\hat{c} = \arg \max_{c=1,\ldots,C} \sum_k I_{\{c\}}(f(k)) * P(k|x) \tag{2}$$

Moreover, the subclasses may be characterized by different variables in the data. If size of training set is not very large, a variable selection may particularly be useful to model only such variables that are relevant to the classification problem.

Example 1

Imagine the case of two classes A and B each consisting of two subclasses A_i and B_i, $i = 1, 2$. Let now the distribution of the subclasses in variable X $f(X|A_i) = f(X|B_i)$, $i = 1, 2$. Figure 1 shows this example for subclasses being normally distributed with unit variance but differing means μ_i. In such case subclasses A_1 and B_2 can be discriminated, as can be subclasses A_2 and B_1. For discrimination of the subclasses A_1 and B_1 as well as A_2 and B_2 this variable contributes no information and should therefore preferably be omitted. This reflection is summarized in the matrix of table 1.

Subclass	A_2	B_1	B_2
A_1	(+)	-	+
A_2		+	-
B_1			(+)

Table 1: $+/-$ indicates whether variable X in example 1 serves for discrimination of two subclasses or not. Parentheses indicate the same (class $c = A$ *or* B). Only half of the subclass-pairs can be discriminated in this variable.

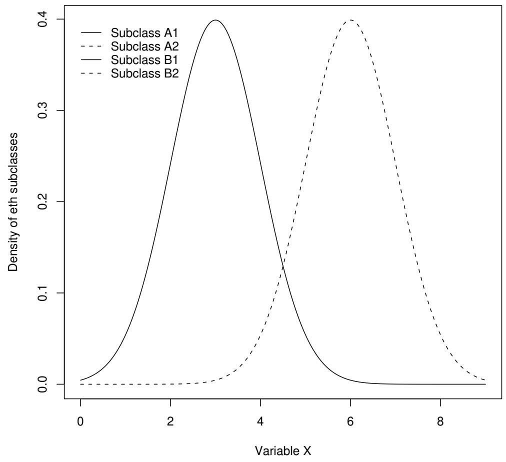

Figure 1: Example of a "2 classes with 2 subclasses each" problem as introduced in example 1. Only half of the subclasses can be separated by differing distributions in this variable.

If any preceeding variable selection in local modelling is desired, this usually has to be performed globally, since comparing local models in different variable subsets is a difficult task. This problem is outlined in Szepannek et al. (2006).
Szepannek and Weihs (2006) proposed a method of pairwise variable selection [PVS]. By this method, the simulated misclassification test-error in the well-known *Waveform* data set (see Breiman et al., 1984) for Linear Discriminant Analysis (which works quite well on this task) has been reduced from 20.02% to 16.96% (being bounded by 14.9% Bayes error from below). A K-class problem is splitted into $K(K-1)/2$ two-class-problems. For any of these class pairs a classification rule is built after some variable selection procedure. The result consists of $K(K-1)/2$ classification models in a "locally maximally reduced" variable space.
Such classification of an observation leads to $K(K-1)/2$ pairwise decisions, returning the same number of pair wise posterior probabilities.

The remaining question consists in building a classification rule from these $K(K-1)/2$ pair wise classifiers.

To solve this task a *Pairwise Coupling* algorithm can be used. It is described in Section 2. If we perform such classification for the subclass-models $k = 1, \ldots, K$ the desired classification can then be obtained by aggregating the subclass-posterior probabilities as in equation 2. This procedure can be performed principally for any classification method returning posterior probabilities in combination with any meaningful method of variable selection.
The following pseudo-code summarizes the steps of the suggested proceeding:

Build.classification.model *(data [containing the subclass.labels], f, classification.method, variable.selection.method)*

f is the function as described above labelling the subclasses to the classes.

1. For each pair of two subclasses do
2. (a) Remove temporarily all observations that do not belong to one of both subclasses from *data*: return *newdata*.
   (b) Perform *variable.selection.method* on *newdata*: return *subspace.of.subclass-pair*.
   (c) Perform *classification.method* on *newdata* only considering *subspace.of.subclass-pair*: return *model.of.subclass-pair*.
   (d) Return *subspace.of.subclass-pair* and *model.of.subclass-pair* for this pair of two subclasses.
3. Return the whole model consisting of: f and for all pairs of subclasses the *subspace.of.subclass-pair* and *model.of.subclass-pair*.

Predict.class *(new.object, subspaces.of.subclass-pairs, models.of.subclass-pairs,f)*

1. For each pair of subclasses do

2. (a) Calculate the class pair wise posterior probabilities for *new object* assuming the object being of in one of the actually considered two subclasses according to *model.of.subclass-pair* on *subspace.of.subclass-pair*.

   (b) Return the *subclass.pair.posterior.probabilities*.

3. Use the Pairwise coupling algorithm to calculate the posterior probabilities for all K subclasses from the set of all estimated pairs of conditional *subclass.pair.posterior.probabilities*, return: *subclass.posterior.probabilities*.

4. Calculate the *class.posterior.probabilities* using the class-labelling function f according to equation 2.

5. Return the predicted class c with maximal *class.posterior.probability*.

The following section describes a solution to the question of gaining the vector of subclass-posterior probabilities form the pair wise classifications built on the different selected variable subsets. Section 3 briefly describes some variable selection methods that are used in the studies in this paper. In Section 4, a simulation study is performed that shows possible benefit of such local variable reduction. In Section 5, the method is applied to some real-world data.

2 Pairwise Coupling

2.1 Definitions

We now tackle the problem of finding posterior probabilities of a K-(sub)class classification problem given the posterior probabilities for all $K(K-1)/2$ pair wise 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 pair wise classification problem to be given by

$$\mu_{ij} = Pr(i|i \cup j) = \frac{p_i}{p_i + p_j} \tag{3}$$

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 2:

Given $p = (0.7, 0.2, 0.1)$. The μ_{ij} can be calculated according to equation 3 and can be presented in a matrix:

$$\{\mu_{ij}\} = \begin{pmatrix} . & 7/9 & 7/8 \\ 2/9 & . & 2/3 \\ 1/8 & 1/3 & . \end{pmatrix} \tag{4}$$

Example 3:

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

$$\{r_{ij}\} = \begin{pmatrix} . & 0.9 & 0.4 \\ 0.1 & . & 0.7 \\ 0.6 & 0.3 & . \end{pmatrix} \tag{5}$$

where the row i contains the estimated conditional pairwise posterior probabilities r_{ij} for class i. 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.

2.2 Algorithm

In this section we present the Pairwise Coupling algorithm of 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:

$$l(\hat{p}) = \sum_{i<j} n_{ij} \left(r_{ij} * log \left(\frac{r_{ij}}{\hat{\mu}_{ij}} \right) + (1 - r_{ij}) * log \left(\frac{1 - r_{ij}}{1 - \hat{\mu}_{ij}} \right) \right) \tag{6}$$

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 details on the IPS-method see e.g. Bishop, Fienberg and Holland, 1975). The algorithm consists of the following three steps:

1. Start with any $\hat{p}$ and calculate all $\hat{\mu}_{ij}$.

2. Repeat until convergence $i = (1, 2, \ldots, K, 1, \ldots)$:

$$\hat{p}_i \leftarrow \hat{p}_i * \frac{\sum_{j \neq i} n_{ij} r_{ij}}{\sum_{j \neq i} n_{ij} \hat{\mu}_{ij}} \tag{7}$$

   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:

Hastie and Tibshirani (1998), show that $l(p)$ increases at each step. For this reason, since it is bounded above by 0, if there exists a proper solution $\hat{p}$ 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 pair wise estimates r_{ij}, the r_{ij} are not independent, but if they were, $l(p)$ of equation 6 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 3, 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 by pairwise coupling algorithms are presented and compared. In the simulations of this paper, the method of Hastie and Tibshirani (1998) is used.

3 Validation of the principle

In this section, the suggested procedure of a subclass pair wise variable selection combined with Pairwise Coupling [PVS] is compared to classification using linear and quadratic Discriminant Analysis [LDA, QDA] with global variable subset selection.

Variable selection:

The method of variable selection in our implementation is a quite simple one. We used subclass pair-wise Kolmogorov-Smirnov tests (see Hajek, 1969, pp.62–69) to check whether the distributions of two subclasses differ in a variable or not. For every subclass pair and every variable, the statistic

$$D = \max_x |F_{n_{k_1}}(x) - F_{n_{k_2}}(x)| \tag{8}$$

is calculated, where the $F_{n_{k_i}}(x)$ are the empirical distributions of subclass k_i, $i = 1, 2$. A variable is taken into a pair wise model if its p value strongly indicates differing densities. Of course, any other variable selection could be used instead. Especially one could refer here to the *stepclass* method (see Weihs et al., 2005) which is a prediction orientated method of variable selection. Variables are included in the model if they improve some predefined measure like e.g. the misclassification rate on the cross-validated data set. This method possesses the advantage that it is adaptive to the specifics of any classification method.

3.1 A first example

Our first example is chosen according to the introducing example 1 in Section 1 to again illustrate the problem. Data are simulated in 3 classes (à 3 subclasses each) and 8 variables. Subclass k is distributed according to $X \sim N(2 * 1.64 * e_k, I)$ if $k < 9$ and $X \sim N(0, I)$, if $k = 9$. Here e_k represents the standard basis vector, 0 is the 0 vector and I is the identity matrix.
This means, two subclasses $k \neq l,\ k, l < 9$ differ in their distributions in only 2 variables $(k\ and\ l)$. Subclass 9 can be discriminated from any other class k only in variable k. Subclasses $k = 1$ to 3 are subclasses of class $c = 1$. Subclasses $k = 4, 5$ and 6 belong to class $c = 2$, so do subclass $k = 7, 8$ and 9 to class $c = 3$. By construction, no variable can be omitted. For that reason, "global" variable selection will not remove any of the variables, using Linear Discriminant Analysis.
Variable selection is especially useful if there are few training examples in the data for estimating the structure of the classes. If classes consist of several subclasses, the amount of available data is further reduced since there are more populations to be fitted with the same amount of data. We therefore computed simulations with varying (equal) (sub)class sizes in the training data to investigate the effect of sparse data. In the test data each subclass contains 50 objects. Error rates are averaged over 50 repetitive simulations of the data set. The results are given in table 2.

size	LDA	QDA	PVS (with LDA)
4	0.186	-	0.154
6	0.140	-	0.110
8	0.123	-	0.096
10	0.112	0.416	0.096
15	0.098	0.240	0.087
20	0.095	0.185	0.086
50	0.084	0.105	0.079

Table 2: Averaged error rates of LDA, QDA and PVS at varying subclass sizes

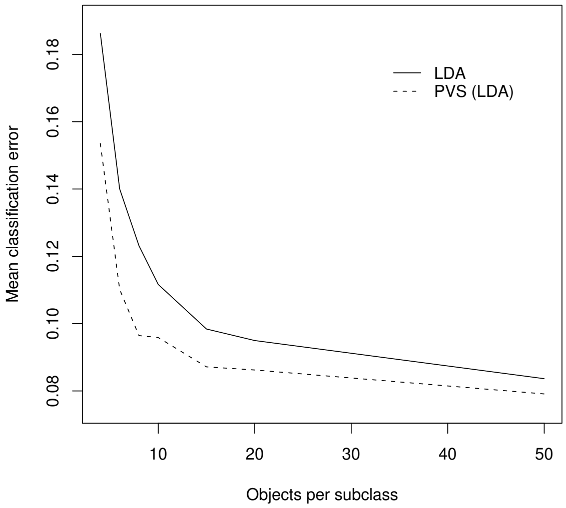

Figure 2: Averaged error rates on test data in simulation 3.1.

The QDA classification rules can only be build having enough data. Even at larger class sizes QDA error rates are still very high. The PVS approach shows systematically lower error rates on the test data than LDA with ”global” variable selection, especially if there are only few observations in the training data. For larger class sizes the differences of both methods in the error rates are still present but seem to vanish.

3.2 Differing variances

We now extend the situation of the first example. In real life it may be possible that one is confronted with data where one of the (sub)classes is strongly concentrated in a specific variable. Of course, this class can be more easily identified by its realizations in this variable. Using LDA will fail to detect this property by pooling all classes' covariances.
We modelled this situation with data consisting of 3 classes each consisting of 3 subclasses (as in the previous example) in 9 variables. Subclass k is distributed following $X \sim N(2e_k, \Sigma)$ with Σ being the identity except from $(\sigma)_{kk} := 0.1$. An illustration of the phenomenon is given in Figure 3 where the vertical line

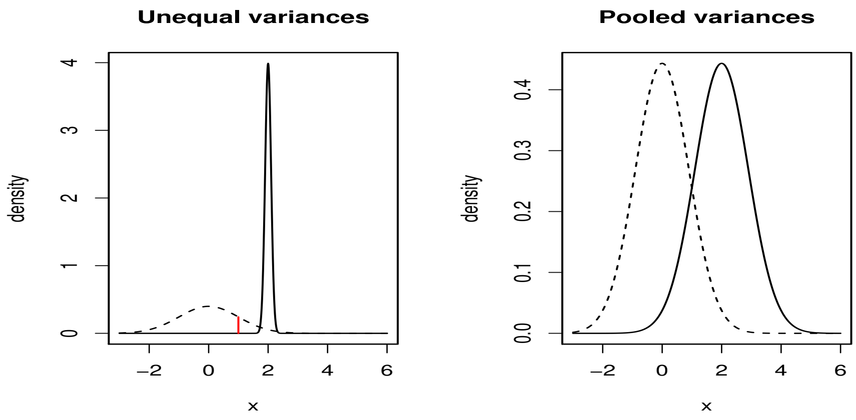

Figure 3: Example of unequal variances and their pooled estimators (by LDA).

in the left plot indicates the wrong 'optimal decision' if wrongly assuming equal covariances as in the right plot. Intuitively, QDA seems to be more appropriate in this situation. The results for varying training data sizes are shown in Figure 4.

size	LDA	QDA	PVS (with QDA)
10	0.250	0.453	0.177
15	0.226	0.273	0.161
20	0.201	0.218	0.151
30	0.182	0.190	0.145
50	0.174	0.171	0.143
100	0.157	0.151	0.133

Table 3: Averaged error rates of LDA, QDA and PVS at varying class sizes

Astonishingly, here LDA still shows smaller error rates than QDA. For QDA, there does not seem to be enough data. Both methods can be largely improved by a class pair wise variable selection using QDA. But note that such variable selection simply using the KS-test statistic will fail to detect situations of correlation between variables.

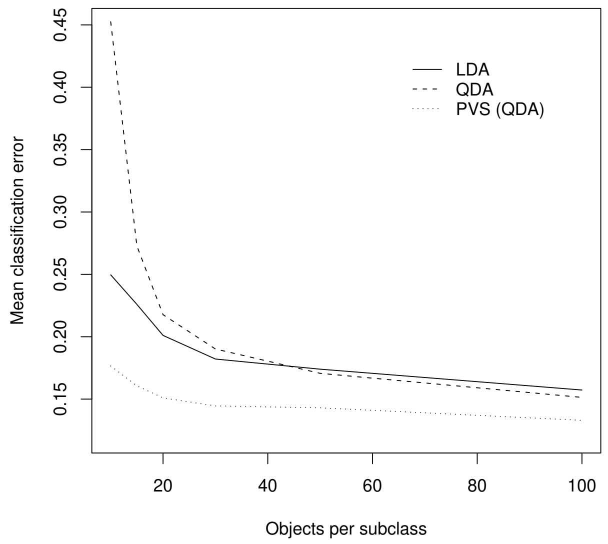

Figure 4: Averaged error rates on test data in example 3.2.

3.3 Real world data

The method is now applied to some real world data. The task is register classification (i.e. correct labelling into high and low pitch) of singers and instruments by pitch-independent features. As predictor variables characteristics of the fundamental and the first 12 harmonics are used. The fundamental $[F0]$ of a sound is exactly its pitch frequency, where the harmonics $[F1, F2, \ldots]$ are all integer multiples of the fundamental frequency. The pitch-independent variables are the mass of the harmonics $F0$ to $F12$ and the width (number of fourier frequencies above some specified threshold in direct neighbourship to the harmonics in the normalized periodogram) without the information about its corresponding frequency.
Figure 5 illustrates the so-called voice print corresponding to the whole song "Tochter Zion" for a particular singer. For masses and widths boxplots are indicating variation over the involved tones (cp. Weihs and Ligges (2003)). For the analyses of this paper we use these characteristics of the voice print for individual tones per harmonic and singer or instrument.
This classification problem may be an example for local modelling as it is de-

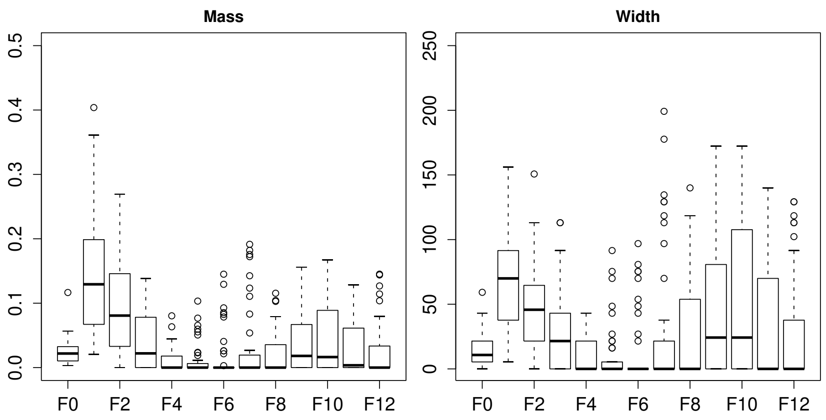

Figure 5: Voice print of professional bass singer.

scribed in the previous sections, since apart from the classes (namely: *high* and *low* register) and the 26 variables also the subclass, i.e. the instrument-type, may influence the distribution of the data. For this reason, local modelling has already been shown here to improve the results.

The data set consists of 432 observations. The subclasses $k := (i, c)$, $i \in \{$*all instruments*$\}$, $c \in \{low, high\}$ are all combinations of instrument i AND register c and contain between 9 and 90 observations. A detailed description of the classification problem as well as a description of the data set and the results of global and local modelling are described in Szepannek et al. (2005). In that paper Linear Discriminant Analysis and Decision Trees are used to build both local and global classification rules. It turned out that the best results are obtained using local LDA-classifiers. Several methods are derived to build classification rules from the local LDA-models for each instrument. The error rates (estimated by leave one out cross validation) have been improved up to 26.9%.

Two of the winner-classification rules are briefly described here:

The first one is referred to as *average density rule*. The estimated multivariate normal densities of the local instrument-subclasses as they are returned by LDA are summed up for the classes, leading to the classification rule:

$$\hat{c} = \arg\max_c \sum_k p(x|k) I_{\{c\}}(f(k)) \tag{9}$$

where $f(k) = f(i, c) = c$ is the function that labels the subclasses $k = (i, c)$ to the corresponding classes c as it is introduced in Section 1 and $p(x|k)$ is the estimated density of the observation given the subclass $k = (i, c)$. Since comparing

densities on different variable subsets is questionable the local models here have to be built on a globally chosen variable subspace.
The second method will be called *global weighting of local posteriors*. It makes use of the fact, that each of the instruments (i.e. the subclasses) appears in combination with all registers in an attribute-like manner and therefore an additional ”global” classification into the correct (unknown) instrument-subclass can be performed. Local LDA classification rules are built for every instrument separately. The obtained local posterior probabilities for the register of a new object are then weighted by some *global weights* that are gained by the posterior probabilities of the ”global” classification into the instrument-subclass. The classification rule can be described by

$$\hat{c} = \arg\max_c \sum_i P(c|i,x) * P(i|x) \tag{10}$$

which is an applicaton of Bayes' theorem. i here denotes the index of the subclass-attribute (instrument). This method turned out to render the smallest obtained error rate. The different local models (given the instrument-subclass) can be built on different variable subsets. But for calculation of the global classification posterior probabilities into the right instrument-subclass of course for all instruments the same variables have to be taken into account.
For comparison, an analysis has been performed using external knowledge about the instrument for the prediction (i.e. an object is classified with respect to the correct local model). Using this extra information the error rates can be improved up to 15% which can be considered as a ”lower bound” for the error rates.
While the *average density rule* does not allow modelling on different variable subsets, the method of *global weighting of local posteriors* does allow models on different feature subsets for different instruments but for the global instrument-classification for all instruments the variables must be the same. For application of this method, it is necessary that the subclasses possess an attribute-like structure. Implementing the PVS method, leads to pairwise comparisons of any combinations (i, c) of instrument and register on possibly differing variables.
Using now the *PVS* approach (with LDA) one observes a further slight improvement of the error rate up to 24.3%. A summary of the different modelling results given in table 4.

method	llo error rate
global LDA	0.345
average density rule	0.301
global weighting of local posteriors	0.269
PVS (with LDA)	**0.243**
”lower bound”	0.150

Table 4: Leave one out cross validated error rates for the different methods

Remark: Relationship between the PVS-method and the ’winner model’

By definition the conditional probability of register, given instrument (and observation x) is given by

$$P(c|i,x) = \frac{P(i,c|x)}{P(i|x)} \tag{11}$$

This changes the classification rule of the ”winner model” of *global weighting of local posteriors* in equation 10 into

$$\hat{c} = \arg\max_c \sum_i P(c,i|x) \tag{12}$$

Using the function $f(k) = f(i,c) = c$ as it is defined above, then our classification rule becomes

$$\hat{c} = \arg\max_c \sum_{(i,c^*)} I_{\{c\}}(f(i,c^*))P(c^*,i|x) \tag{13}$$

This classification rule is of the same form as it is introduced in equation 2 in Section 1 for local modelling by the PVS approach. It can be seen, that in both methods modelling is essentially done in the same way. The difference is in estimating the local membership probabilities. The PVS method here only uses those variables that are important for decision between two subclasses. This explains why the result of the winner rule is even slightly improved by using the proposed method.

Additionally, the proposed PVS method is more flexible since it can also be applied to subclasses that do not possess an attribute-like character as the subclasses in the example do.

4 Summary

The problem is tackled to perform local modelling for classification where the variable subspaces of the different local models can differ. An approach of pair-wise variable selection [PVS] is suggested to perform the maximal possible variable selection by splitting a K-subclass classification problem into $K(K-1)/2$ subclass pair-wise classification problems. An algorithm is presented to build a classification rule from the results using this method. This principle can be applied to any classification method returning class-membership posterior probabilities in combination with any (meaningful) variable selection procedure.
Situations are outlined where such proceeding is strongly beneficial. The method is investigated on different simulated and real world data sets using (linear and quadratic) Discriminant Analysis and the results are compared to their original results using global variable selection. Gain in classification error rate can be noticed, especially if the number of observations is not very large.
Additionally, the pairwise variable subset selection can give interpretational insight into which features characterize the differences between two (sub)classes.
On the other hand, the computation time grows since there have to be built $K(K-1)/2$ classification models. Furthermore, the classification rule of each object has to be iteratively evaluated by the Pairwise Coupling algorithm.

Acknowledgment. This work has been supported by the Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475) of the German Research Foundation (DFG).

References

BISHOP, Y., FIENBERG, S. and HOLLAND, P. (1975): Discrete multivariate analysis, *MIT Press, Cambridge*.

BRADLEY, R. and TERRY, M. (1952): The rank analysis of incomplete block designs, i. the method of paired comparisons, *Bimometrics, 324–345*.

BREIMAN, L. FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984): Classification and regression trees. *Chapman & Hall, NY*.

HAJEK, J. (1969): A course in nonparametric statistics. *Holden Day, San Francisco*.

HASTIE, T. and TIBSHIRANI, R. (1998): Classification by Pairwise Coupling. *Annals of Statistics, 26(1), 451–471.*

SZEPANNEK, G., LIGGES, U., LUEBKE, K., RAABE, N. and WEIHS, C. (2005): Local Models in Register Classification by Timbre. *Technical Report 47/2005, SFB 475, Fachbereich Statistik, Universität Dortmund.*

SZEPANNEK, G. and WEIHS, C. (2006): Variable Selection for Discrimination of moren than two Classes where Data are Sparse. *In: M. Spiliopoulou, R. Kruse, A. Nürnberger, C. Borgelt, W. Gaul (Eds.): From Data and Information Analysis to Knowledge Engineering, Springer-Verlag, Heidelberg (accepted).*

WEIHS, C. and LIGGES, U. (2003): Voice Prints as a Tool for Automatic Classification of Vocal Performance. In: R. Kopiez, A.C. Lehmann, I. Wolther and C. Wolf (Eds.): *Proceedings of the 5th Triennial ESCOM Conference*. Hannover University of Music and Drama, Germany, 8-13 September 2003, 332–335.

WEIHS, C., LIGGES, U., LUEBKE, K. and RAABE, N. (2005): klaR Analyzing German Business Cycles. In: D. Baier, R. Becker and L. Schmidt-Thieme (Eds.): *Data Analysis and Decision Support*, Springer, Berlin, 335–343.

WU, T.-F., LIN, C.-J. and WENG, R. (2004): Probability Estimates for Multi-class Classification by Pairwise Coupling. *Journal of Machine Learning Research, 5, 975–1005.*