Understanding Patterns with Different Subspace Classification

Gero Szepannek, Karsten Luebke and Claus Weihs *

Department of Statistics
University of Dortmund
szepannek@statistik.uni-dortmund.de

Abstract. By identifying *characteristic regions* in which classes are dense and also relevant for discrimination a new, intuitive classification method is set up. This method enables a visualized result so the user is provided with an insight into the data with respect to discrimination for an easy interpretation. Additionally, it outperforms Decision trees in a lot of situations and is robust against outliers and missing values.

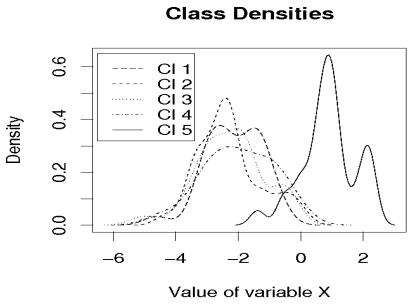
1 Introduction

Classification or supervised learning often involves two goals: the first is allocation or prediction, i.e. assigning class labels to new observations. The second goal, which can be even more important, is descriptive and involves the discovery of the underlying differences between the classes. The new Different Subspace Classification (DiSCo) method is a method to simultaneously visualize and classify multi-class problems in high dimensional spaces and is therefore designed to attain both predictive and descriptive goals.
Decision trees and Naive Bayes classifiers are two of the most often used data mining techniques. In case of Decision trees this may be due to the fact that the result of a tree can often be interpreted in terms of the subject matter (see e.g. Hastie et al. 2001, p. 267). Furthermore, Decision trees perform variable selection: variables which are not relevant for classification are not used to build the tree. A shortcoming of trees is that in the final tree only parts of the marginal distribution of the variables are used, conditional on the split. Another major problem caused by the hierarchical structure of a tree is the inherent instability to small changes in the data resulting in high variance (Hastie et al. 2001, p.274). To overcome this, Random Forests (Breiman 2001) and Bagging (Breiman 1996) can be applied but then the easy interpretation is lost. The Naive Bayes method is somewhat different. There, all class-conditional univariate marginal densities are estimated independently. Especially in high dimensional feature spaces Naive Bayes often performs well (see e.g. Hastie et al. 2001, p. 185). Unfortunately the result of Naive Bayes is not so easy to interprete and it can not be used to select variables. Also it is not robust against outliers.

* This work has been supported by the Collaborative Research Center 475 of the German Research Foundation (DFG).

The higher the dimension of the data the more challenging is the understanding of the data. So if there are many observed variables, methods of variable selection are often used to reduce the dimension of the data. Such methods identify and retain those of the variables that separate the classes best – like a Decision tree does. Afterwards, a classification method is (re-)applied to the resulting subspace of variables. A problem may be that in general the variables do not contain relevant separating-information for all classes. So, a variable can contain information for separating class i from the rest but no information for the discrimination of class $j \neq i$. This may be illustrated by Figure 1. Density estimation of five classes is shown and it can be seen that by this variable, an object of class 5 may be probably well separated from the others. But a value of e.g. -2 will not tell us much about its real class (which may be probably one of the classes 1 to 4). The new DiSCo method can be considered as a mixture

Fig. 1. Density estimation of five classes.



of both Decision trees and Naive Bayes: calculate all class-conditional univariate marginal densities by an appropriate kind of histogram estimate comparable to Naive Bayes and find out the so called *dense* regions, where many objects of a class fall in. Check whether these regions are *relevant* to distinguish this class from others and take away all dispensable information like a Decision tree. Both ideas together are used in the new method to tackle classification problems. Moreover, it can be seen that it is robust against outliers, missing values and can be used with metric and categorical data. In DiSCo variable selection is intrinsic to the classification method. The resulting subsets of variables which are used for discrimination of the classes can differ between the classes. Another focus of the new classification method lies on the visualization of the class-characteristics. The proposed method does not make any assumptions about the underlying distribution of the data. The only, not very strong assumption is that objects of the same class are similar in some of their observed variables.

In the following section the concept of *characteristic regions* is defined and a classification rule is developed. Section 3 explains the visualization of the results. Section 4 briefly summarizes the choice of parameters for the implementation of the method while section 5 contains results of a comparative study of the three mentioned methods on simulated data.

2 Notation and Method

The idea of the new method is to search for characteristic regions, i.e. sets of values in some variables that indicate the class-membership. To build up these characteristic regions two steps are needed. The first step is to search for intervals of the realizations of the random variables that contain a large probability mass of the classes. The resulting "regions" are called dense regions. The second step, which is independent of the first, identifies regions that discriminate at least one class from the others because of a relatively high density. These regions are called relevant regions. Regions that are both dense and relevant are then called characteristic regions.

2.1 Characteristic Regions

The concept of the characteristic regions is given as follows:

Definition 1.

- *For metric variables X^d (where d is the variable index):*
  S^d being the set of all possible realizations of an object x_n in variable X^d, for each d let $\{R_m^d : 0 \leq m \leq M^d + 1\}$ be a contiguous segmentation of an interval covering S^d following
  1. $\bigcup_{m=0}^{M^d+1} R_m^d \supseteq S^d$
     (All possible realizations of X^d are covered by the union of all its regions.)
  2. $\forall\, x_1, x_2 \in R_m^d$ *and* $\alpha \in [0,1]$: $\alpha x_1 + (1-\alpha)x_2 \in R_m^d$
     (The regions of every variable are contiguous.)
  3. $\forall\, x_1 \in R_{m_1}^d, x_2 \in R_{m_2}^d,\ m_1 < m_2$: $x_1 < x_2$
     (In every variable the regions are disjoint and also ordered.)

  *R_m^d are called **regions** of variable X^d. $M^d(+2)$ denotes the number of regions variable X^d. A possible choice is proposed in section 4.1.*
  By restriction 2 all the objects that fall into one region can be considered to be similar.
- *For categorical variables X^d:*
  If a variable is categorical the **regions** *are implicitly given by all its possible values. Sometimes it may be reasonable for the user to merge some of the values to one region if there are too many levels or because of the subject matter.*

Definition 2. *Let x_n^d be the value taken by object n in variable X^d and let k_n be the corresponding, known index of its class. Then*

$$n_m^d(k) := \sum_{n=1}^{N} I_{[R_m^d]}(x_n^d)\, I_{[k]}(k_n) \tag{1}$$

*with $I_{[\cdot]}$ as the indicator function is called the **corresponding frequency** of class k in Region m of variable d.*

As the $n_m^d(k)$ should represent the density of the data it is assumed for simplicity of comparisons that for any fixed d and all $1 \leq m \leq M^d$:

$$\sup_{x \in R_m^d} - \inf_{x \in R_m^d} \equiv const.,$$

so the regions of a variable have equal width. $m = \{0, M^d + 1\}$ are necessary to form "outer regions" (see section 4.1). By this the corresponding frequencies are proportional to heights of histogram bars of the classes if the bandwidths are given by the regions.

Let ***dense regions*** be those regions which contain most of the classes' probability masses. Let $S_{DR} > 0$ be a threshold to construct class wise dense regions. Then, dense regions are regions $R_{m_0}^d(k)$ with

$$n_{m_0}^d(k) \geq S_{DR} \frac{\sum_{m=0}^{M^d+1} n_m^d(k)}{M^d} \tag{2}$$

This proceeding corresponds to comparing the observed corresponding frequency to the mean over all regions.

Relevant regions should be the regions where the density of one class k is high compared to those of the other classes and so a new observed object lying in this region strongly indicates its membership to class k. Let $S_{RR} > 0$ be a threshold to construct class-wise relevant regions. Then, relevant regions are regions $R_m^d(k_0)$ with:

$$\frac{n_m^d(k_0)}{N_{k_0}} \geq S_{RR} \frac{\sum_{k=1}^{K} \frac{n_m^d(k)}{N_k}}{K} \tag{3}$$

with K being the number of different classes. To be able to compare the regions' densities of different classes by corresponding frequencies they have to be weighted by their observed absolute frequencies N_k. Finally, ***characteristic regions*** are regions that are both dense and relevant.
Missing values in one or more variables can simply be omitted when building the (variable-wise) regions without loss of information for the other variables.

2.2 Classification Rule

Let $w_m^d(k) \geq 0$ be a ***class wise weight of a region*** of class k connected to region R_m^d.
The characteristic regions are used to build up the classification rule by summing the weights over all variables. Then the assignment of the class is obtained by

$$\hat{k}(x_{new}) = \arg\max_k \sum_{d=1}^{D} \sum_{m=0}^{M^d+1} I_{[R_m^d]}(x_{new}) w_m^d(k) \tag{4}$$

where the weights of the characteristic regions are defined by

$$w_m^d(k_0) := \begin{cases} 0 & \text{if (2) or (3) do not hold} \\ \frac{n_m^d(k_0) \frac{p(k_0)N}{N_{k_0}}}{\sum_{k=1}^{K} n_m^d(k) \frac{p(k)N}{N_k}} & \text{if } R_m^d \text{ is characteristic for class } k_0 \end{cases} \tag{5}$$

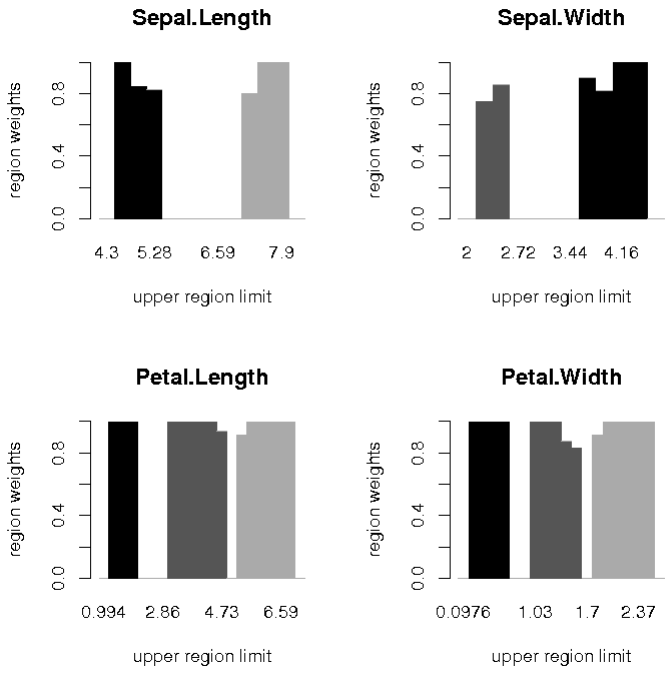
where $\frac{p(k)N}{N_k}$ adjusts the *corresponding frequencies* if the observed class frequencies differ from known a priori class probabilities $p(k)$. The weights are motivated by the marginal probability of $k_{new} = k$ given $x^d_{new} \in R^d_m$, if R^d_m is "characteristic" for class k.

3 Visualization

The weights $w^d_m(k)$ described above mimic marginal conditional probability of the different classes. As only characteristic regions will be shown in our visualization only robust information relevant for classification is given. So plotting these class wise weights of the regions (see equation 5) provides a visualization of the class characteristics and an interpretation may be simplified.

As example we illustrate the method in Figure 2 on the well known Iris data set.The values of the variables are shown on the x-axes while the different colours of the bars symbolize the different true classes (black = "Setosa", light grey = "Virginica" and dark grey = "Versicolor"). The heights are the weights of the characteristic regions. It can be seen that the variable "Sepal length" only

Fig. 2. Example: Visualization of a result for Iris data



serves to indicate membership of one of the classes "Virginica" or "Setosa" but

not for "Versicolor", while the variable "Sepal width" just serves to characterize a plant of class "Setosa" or "Versicolor". The "Petal" variables seem to separate all three classes with the lowest values for class "Setosa". The upper extreme values indicate the class "Virginica". As the plots of these two variables are of the same structure one can suppose a correlation between these variables.

4 Implementation of the Method

4.1 Building the Regions for Metric Variables

As mentioned earlier the *corresponding frequencies* are proportional to heights of histogram bars for simplicity, so we can refer to the theory of nonparametric density estimation to build the regions. In histogram density estimation a problem consists in smoothing but not over-smoothing the empirical distribution of the data. Thus the bin-width of a histogram should be chosen neither too small nor too large. Freedman et al. 1981 suggest a choice of

$$bw = \frac{2}{\sqrt[3]{N}}\, IQR \tag{6}$$

as bin-width where IQR is the interquartile range. Under weak assumptions this histogram is L^2-convergent for density estimation (Freedman et al. 1981). As the distribution may be different in the classes this bin-width calculation must be done for every class and every variable separately, returning $bw(k,d)$. The number of class-wise bins is then $M^d(k) = r(\frac{x^d_{(N_k)} - x^d_{(1_k)}}{bw(k,d)})$ with $x^d_{(N_k)}$ and $x^d_{(1_k)}$ being the class-wise maximum or minimum, respectively, $r(\cdot)$ being the rounding operator. With $IV^d := [x^d_{(1)}, x^d_{(N)}]$ and $IV^d_k := [x^d_{(1_k)}, x^d_{(N_k)}]$ let:

$$M^d := r\left(\left\{\sum\nolimits_k \left(M^d(k) \int_{IV^d_k} \{\sum\nolimits_k I_{[IV^d_k]}(s)\}^{-1} ds\right)\right\} * \frac{\int_{IV^d} 1dt}{\int_{\cup_k IV^d_k} 1dt}\right) \tag{7}$$

In the simplest case, if the densities of the classes do not overlap and also there is no interspace between them, (7) reduces to $M^d = \sum_k M^d(k)$. If there is space between the classes this space must be filled also with bins. Therefore, in (7) with $\int_{IV^d} 1dt$, the width of the whole interval IV^d, is related to $\int_{\cup_k IV^d_k} 1dt$, the width of those parts of the whole interval which are covered by values of the different classes. If there is some free space between the classes this is smaller than the whole interval and the number of bins is linearly projected. In cases when densities of the classes do overlap this must also be corrected. By the calculation of $\{\sum_k I_{[IV^d_k]}(s)\}^{-1}$ it is assured that those parts of IV^d_k which are covered by more classes than just class k are not repeatedly counted in the calculation of M^d. So the class wise number of bins is linearly projected or averaged, respectively, for intervals covered by $0, 1, 2, \ldots$ classes. In other words M^d is a linear projection of the k class-wise number of bins $M^d(k)$ of the parts covered by the different classes to the whole range of the data averaged for the

classes' overlap. The regions of variable d are IV^d divided into M^d equal parts. R_0^d and $R_{M^d+1}^d$ cover the upper and lower rest.

By construction DiSCo is robust to outliers as outliers are not dense by definition. The only problem that may occur is that an unnecessarily high value for the number of bins is calculated. Therefore class wise outliers can be removed in the calculation of the number of bins without any loss of information. Outliers can be for example all observations that differ more than a fixed value (for example $\frac{3}{2}$ times the standard deviation) from the variables' median.

4.2 Optimizing the Thresholds

There remains the question how to choose the thresholds in equation 2 and equation 3. So far no theoretical background is known for an optimal choice of both S_{DR} (dense regions) and S_{RR} (relevant regions). The optimal parameters are found by a 2-dimensional grid-search algorithm. As the criterion for optimization the cross validated error rate (on training data) is used. Concerning the parameters one can suppose that a rather small threshold S_{DR} eliminates outliers but keeps a large probability mass in the remaining regions. S_{RR} rather large keeps only regions in the model that strongly indicate one class.

5 Benchmark Study

In the previous sections, we focussed on classification methods that work on the variables separately, namely Classification trees, Naive Bayes and the newly developed DiSCo method. We will now compare these methods in a quite general simulation study to investigate the advantages of each method.

We will start with the simple case of normally distributed data in the following subsection. Then, we turn to situations where the assumption of normality is violated. Subsection 5.2 simulates situations with multimodality in the data and in subsection 5.3 the effect of outliers is investigated.

5.1 Normally Distributed Data

We simulated data consisting of three classes and three variables. Each class is separated from the other classes by a different mean in one variable – while the other two classes have the same mean in that variable (compare Figure 1 where the mean of class 5 is separated from classes 1-4). All variables are normally distributed with variance 1. The location difference is chosen to be twice the α-quantile of the standard normal distribution, guaranteeing a controlled probability of overlap and therefore misclassification of the classes. So the Bayes risk for the separated class is α. For the other two classes with same means the expected error rate equals 0.5 so that a random choice will be as good. All simulations rely on 300 objects in both training and test data set each class having the same prior probability. The results are averaged over all 30 repeated simulations.

Table 1 shows the effect of the classes' overlap in normally distributed data on the misclassification rate for the different methods. Of course, since the assumption of normality holds Naive Bayes turns out to have lowest error rates. For large location differences (i.e. small overlapping probabilities) all methods show very small error rates, as expected. In such situations, DiSCo is preferable to Classification trees since the error rates of the trees are up to 50% higher compared to those of DiSCo. We also tested deviation from normality by different skewness levels, but this had almost no influence the performance of the methods compared to each other so these results are omitted here.

Table 1. Test error rates on normally distributed data at varying probabilities of class-overlap

Class overlap	Naive Bayes	CART	DiSCo
0.010	0.001	0.021	0.016
0.050	0.017	0.065	0.043
0.100	0.063	0.109	0.099
0.400	0.525	0.572	0.578

5.2 Effect of Multimodality

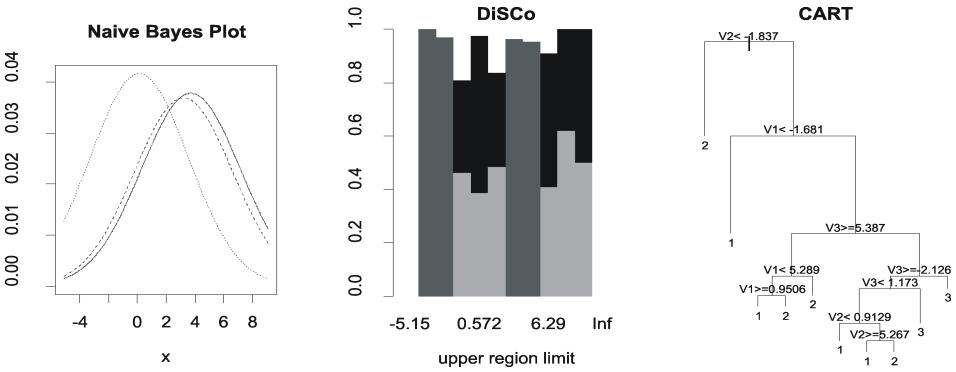
Another violation of normality may be caused by multimodality of the data. This seems to be an important case for practical applications since classes may consist of several different "subclasses", leading to multimodal distributions. We constructed data as in section 5.1 but with each class possessing a bimodal distribution. The distributions are designed as follows: an object is with probability $p = 0.5$ from one of two normal distributions $N(0, 1)$ or $N(2*\alpha, 1)$. In each of the three variables two of the classes are identically distributed following the bimodal distribution specified before. The third class differs in location to both others in a manner that the two underlying distributions are shifted to be $N(-\alpha, 1)$ or $N(\alpha, 1)$. α is varied to investigate different levels of overlap of the classes. It determines the overlap of two neighbouring modes and is varied as in section 5.1. The results (Table 2) show unacceptably large error rates when wrongly assuming (unimodal) normally distributed data as for the Naive Bayes method. The DiSCo error rates dominate those of the Classification tree. With increasing overlap the performance of Naive Bayes is approximating those of the the other methods. The visualization of the results of the three different methods is shown in Figure 3. The Decision tree visualizes the whole decision rule and is therefore maybe the most comprehensive way to display the entire decision. Nevertheless, since there are many conditional splits in the tree, the specific characteristics of the three classes are hardly identifiable. The results of Naive Bayes and the DiSCo method can be visualized for each variable separately. For Naive Bayes,

Table 2. Test error rates on bimodal data at varying overlap percentages between two neighbouring densities

% overlap	Naive Bayes	CART	DiSCo
0.001	0.379	0.010	0.004
0.050	0.386	0.125	0.094
0.100	0.406	0.232	0.192
0.200	0.432	0.416	0.401

one can plot the density estimation of the different classes. This here gives a completely wrong impression of the structure of the classes, while DiSCo (as introduced in section 3) nicely displays the classes' characteristics: all modes are identified and it can be seen whether their locations are characteristic for one single or more than one class.

Fig. 3. Visualization of the results of the different methods on bimodally distributed data, only one variable is shown for Naive Bayes and DiSCo.



5.3 Effect of Outliers

The last study investigates the behavior of the three methods if the data are contaminated with outliers. The data sets are generated as in section 5.1 except that with chance of 5% an object is an outlier following a distribution $N(0, \sigma^2)$ with variance much larger than those of the classes' distributions. The classification errors are examined for different sizes of σ.
Classification trees and DiSCo behave relatively robust to outlier contamination while Naive Bayes becomes much worse with increasing variance of the outliers.

We further observe that the misclassification rates of Classification trees are systematically worse than those of DiSCo.

Table 3. Test error rates on outlier contaminated data at varying variance of the outlier distribution

SD of outliers	Naive Bayes	CART	DiSCo
3	0.032	0.070	0.054
10	0.052	0.079	0.057
20	0.099	0.077	0.059
50	0.331	0.078	0.067

6 Summary

Motivated by the fact that different regions in the variable may discriminate some but not all classes a new classification method is set up. By identifying *characteristic regions* that indicate whether regions of values are dense and also relevant for discrimination this method implicitly includes a feature selection. Moreover, it is robust to outliers and missing values in the observed data. Also the descriptive aspect of data analysis is addressed by an informative visualization of the DiSCo result.

A benchmark study is performed where the new method is compared to Classification trees and the Naive Bayes classifier since both methods also work on the marginal data. Different situations are examined. Comparing the missclassification rates, the Naive Bayes classifier performs better than both other classifiers if the assumption of normality holds while DiSCo has smaller error rates than the Classification tree. If the data are generated from multimodal distributions or contaminated with outliers, Naive Bayes' error rates become unacceptably high. The other two methods are able to handle such data while the misclassification rates of the Classification trees are slightly dominated by those of DiSCo.

References

Breiman, L.: Bagging predictors. Machine Learning **26** (1996) 123–140

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

Breiman, L., Friedman, J., Olshen, R. and Stone, C.: Classification and regression trees. Wadsworth Publishing Co Inc. 1984.

Freedman, D. and Diaconis, P.: On the histogram as a density estimator: L_2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **67** (1981) 453–476

Hastie, T., Tibshirani, R. and Friedman, J.: The elements of statistical learning. Springer. 2001.