Different Subspace Classification

Gero Szepannek * and Karsten Luebke

University of Dortmund**, Department of Statistics, 44221 Dortmund, Germany

Abstract. We introduce the idea of **Characteristic Regions** to solve a classification problem. By identifying regions in which classes are dense (i.e. many observations) and also relevant (for discrimination) we can characterize the different classes. These Characteristic Regions are used to generate a classification rule. The result can be visualized so the user is provided with an insight into data for an easy interpretation.

1 Introduction

Supervised Classification or Discrimination 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 disclosure 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 therefore is designed to attain both predictive and descriptive goals.

The problem of classification or pattern recognition is given in the following way: N objects $x_n, n = 1, \ldots, N$, are observed, each object belonging to one and only one class $k_n, k_n \in \{1, \ldots, K\}, n = 1, \ldots, N$. The class membership is known to the user. N_k objects are observed from class k. This set of objects is called training data. For each object D variables $x^d, d = 1, \ldots, D$, are observed. Every object x_n can be considered as a D-dimensional realization of a random vector X_n following an unknown distribution that depends on its class k_n.

The first goal is to be able to determine the correct (unknown) class for objects x_{new} that will be observed in future. The second goal is to find out the characteristics of the different classes by analyzing the training data. 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. These methods identify and retain those of the variables that separate the classes best. Then following this procedure a classification method is (re-)applied to the resulting subspace of variables. A problem may be that in general the variables do not contain equal separating-information for all classes. So a variable can contain

** This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

* e-mail: szepannek@statistik.uni-dortmund.de

information for separating class i from the rest but no information for the separation of class $j \neq i$.

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.
A focus is also laid on the visualization of the class-characteristics. The proposed method does not make any assumptions about the underlying distribution of the data. The only weak assumption is that objects of the same class are similar in some of their predictor values.

In the following chapter the principle of Characteristic Regions is defined and a classification rule developed. Chapter 3 explains the visualization of the results. Chapter 4 briefly summarizes the choice of parameters for the implementation of the method while chapter 5 contains a simulation study with comparison to Classification trees and Discriminant analysis.

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

Definition 1. S being the set of all possible predictor values of an object x_n, for all d let $\{R_m^d \ : 0 \leq m \leq M^d + 1\}$ be a contiguous segmentation of an interval covering $S \cap X^d$ following

1. $\bigcup_{m=0}^{M^d+1} R_m^d \supseteq S \cap X^d$
   (All possible values 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.

By restriction 2 all the objects that fall into one region can be considered to be similar.

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. 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 classwise 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 classwise Relevant Regions. Then Relevant Regions are regions $R_m^d(k_0)$ with:

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

To be able to compare the regions' densities of different classes by corresponding frequencies they have to be weighted by their observed absolute frequencies. Finally, ***Characteristic Regions*** are regions that are both dense and relevant.

2.2 Classification Rule

Let $w_m^d(k) \geq 0$ be the ***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

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

$\frac{p(k)N}{N_k}$ is a correction term for the absolute frequency of the classes in the data with the prior probabilities of the classes – if it differs from the observed frequency. The weights are motivated by the marginal probability of $k_{new} = k$ given $x_{new}^d \in R_m^d$, if R_m^d is "characteristic" for class k.

As only Characteristic Regions are used for the classification rule the cutpoints of the regions may disregard information. So to keep more of the classes' probability masses we propose another smoothed classification rule where the weights $w_m^d(k)$ are as before but additionally the adjoining regions are included in the model. Then:

$$\hat{k}(x_{new}) = \arg\max_k \sum_{d=1}^{D} \sum_{m=0}^{M^d+1} I_{[R_m^d]}(x_{new})(\frac{1}{2}w_{m-1}^d(k) + w_m^d(k) + \frac{1}{2}w_{m+1}^d(k)) \tag{6}$$

with $w_m^d(k) = 0$ for $m = -1, M^d + 2$.

3 Visualization

The weights $w_m^d(k)$ 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 1 on the well known Iris data set introduced by Fisher. 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 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.

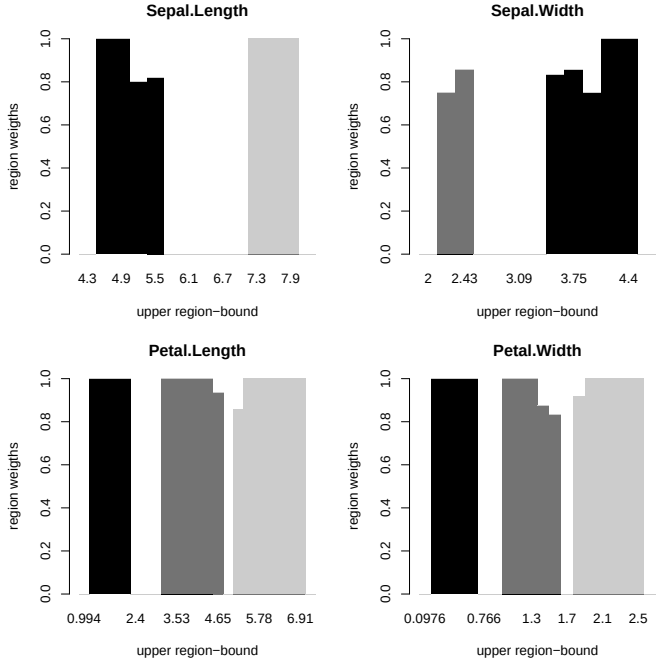

Fig. 1. Example: Visualization of a result for Iris data

4 Parameter choice for DiSCo

For the implementation one has to find the Characteristic Regions. So the problem is how to form the regions and how to choose the thresholds.

4.1 Building the regions

As mentioned earlier the corresponding frequencies should be proportional to heights of histogram bars for convenience so we can refer to the theory of nonparametric density estimation to build the regions. In histogram density estimation the problem consists in smoothing but not over-smoothing the empirical distribution of the data. Thus the bandwidth of a histogram should be chosen neither too small nor too large. Freedman and Diaconis (1981) suggest a choice of

$$bw = \frac{2}{\sqrt[3]{N}}\, IQR \tag{7}$$

as bandwidth where IQR is the interquartile range. Under weak assumptions this histogram is L^2-convergent for density estimation (Freedman and Diaconis, 1981). As the distribution may be different in the classes this must

be done for every class – and every variable. The number of classwise bins is then $M^d(k) = \lfloor \frac{x^d_{(N_k)} - x^d_{(1_k)}}{bw(k,d)} \rfloor$ with $x^d_{(N_k)}$ and $x^d_{(1_k)}$ being the classwise maximum respective minimum and $\lfloor . \rfloor$ 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:

$$M^d := \left\lfloor \frac{\int_{IV^d} 1dt}{\int_{\cup_k IV^d_k} 1dt} \left\{ \sum\nolimits_k \left(M^d(k) \int_{IV^d_k} \{ \sum\nolimits_k I_{[IV^d_k]}(s) \}^{-1} ds \right) \right\} \right\rfloor \tag{8}$$

This means that the classwise number of bins is interpolated resp. averaged for intervals covered by none, one or more than one class. So the regions of variable d are IV^d divided into M^d equal parts. R^d_0 and $R^d_{M^d+1}$ cover the upper and lower rest.

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 contracting 2-dimensional grid-search algorithm. As the criterion for optimization the cross validated error rate is used. It should be noticed that since the number of observations is finite small changes of the two thresholds will not change the resulting model. In order to check the parameters one can consider that a rather small threshold S_{DR} eliminates outliers but keeps a large probability mass in the remaining regions. A S_{RR} rather large keeps only regions in the model that strongly indicate one class.

5 Simulation study

5.1 Data generation

In order to obtain more general results an experimental design is used in data generation to be able to compare the effects of possibly influencing factors in the data on the classification result of DiSCo and of two well-established other methods: Classification Trees (CART) and Linear Discriminant Analysis (LDA).

With the factor levels described below, data of 8 or 12 variables are first drawn from independent normal distributions with variance 1 but different expectations in 3 classes. These data are transformed to possess different kurtosis and skewness and to be deflected.

Below we give a brief description of the seven investigated factors:

- The class priors may be equal or not.

- We investigated two different class mean settings in the first 6 variables: either only one class mean separated from the others or all three class means are different (one in the middle between the others). For 3 variables the doubled 0.95 quantile is chosen, for the 3 other variables the doubled 0.9 quantile of the standard normal distribution is chosen for the tallest differences in the location of the class means.
- 2 or 6 irrelevant independent variables are attached to the data that are $N(0, 1)$ distributed for all classes, i.e. either a quarter or half of the variables do not contain any separating information.
- All variables are transformed to have high or low kurtosis and skewness following the Johnson-System (see Johnson, 1949) to generate a wide range of values in the kurtosis-skewness-plane.
- The probability of an object to be deflected is fixed to be 0.1 or 0.4, where deflection means that an object is moved into one of two directions: half the distance towards its class mean or half the way away from it into the direction of its nearest wrong class.

The factors and levels included in the experimental design of the simulation study are summarized in Table 1. A Plackett-Burman design (Plackett and Burman, 1946) for these factors was repeated 20 times.

Effect	Low level	High level
Class priors	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$
Number of different class means	3	2
Added irrelevant variables	2	6
Kurtosis	2.7	5
Skewness	0.1^2	1.15^2
Probability to be deflected	0.1	0.4
Direction of deflection	towards class mean	away from class mean

Table 1. Effects and levels on the simulated data sets

5.2 Results

Compared are both proposed classification rules for the DiSCo method including (labelled (1)) and not including (2) the adjoining regions, CART (Breiman et al., 1984) and LDA. Table 2 shows the mean error rates on the test data and the estimated effects of the main factors (coded to $-1/+1$) used in the design (cp. table 1) on $log(odds(hitrate))$. These effects can be estimated independently by a regression on the coded influencing factors:

DiSCo seems to outperform the Classification trees and is almost as good as LDA. One can also see that there are only small differences between both proposed classification rules for the DiSCo method so there is no general rule which one to use.

	DiSCo (1)	DiSCo (2)	CART	LDA
Overall mean error	0.085	0.079	0.127	0.075
Class Priors	0.41	0.48	0.19	0.24
Number of different class means	0.17	0.24	0.34	0.37
Irrelevant variables	0.19	0.26	0.21	-0.11
Kurtosis	-0.12	-0.07	-0.27	0.09
Skewness	1.16	1.06	0.89	0.70
Probability to be deflected	-0.10	-0.16	-0.24	-0.61
Deflected direction	-2.17	-2.23	-1.00	-2.73

Table 2. Results: Overall mean error and estimated effects on $log(odds(hitrate))$

It can be concluded that LDA has best overall mean error. Classification trees perform well with deflection away from the class mean but having a large general deficit. The DiSCo method, having a good average result, is preferable with skewed data or differing class priors and a high percentage of deflected objects.

Mean values for the optimal thresholds are $S_{DR} = 0.67$ and 0.54 including and not including the neighbour regions while the averaged optimal S_{RR} are 1.88 and 1.75.

6 Summary

The introduced concept of **Characteristic Regions** allows the visualization of the class characteristics and so satisfies the aim of an easy comprehension and interpretation of the data. It also yields intuitive classification rules.
On simulated test data it outperformed classification trees and was almost as good as the linear discriminant analysis.

References

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984): *Classification and regression trees*. Chapman & Hall.

Freedman, D. and Diaconis, P. (1981): On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitsthoerie und verw. Gebiete, 57, 453-476.*

Johnson, N. L. (1949): Bivariate distributions based on simple translation systems. *Biometrika, 36, 149-176.*

Plackett, R., Burman, J. (1946): The design of optimum multifactorial experiments. *Biometrika, 33, 305-325.*