ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

Daniel Kosiorowski[*]

ROBUSTNESS OF DEPTH BASED CLASSIFICATION RULES

Abstract. In this paper we propose several classification rules based on data depth concept. We study a performance of the propositions on various multivariate data sets simulated from skewed, fat tailed distributions and mixtures of them. We discuss also a rule allowing for choosing a correct number of classes $A_1, \ldots, A_k$ portioning the data set A.

Key words: Robust statistical procedure, discriminant rule, statistical depth function.

I. INTRODUCTION

Consider k p-dimensional populations $C_1, \ldots, C_k$, $k \geq 2$. Suppose that associated with each population C_j, there is a probability density $f_j(\mathbf{x})$ on $\mathbb{R}^p$, so that if an individual comes from population C_j, he has p.d.f. $f_j(\mathbf{x})$. Then the object of discriminant analysis is to rationally allocate an individual to one of these populations on the basis of his measurements $\mathbf{x} \in \mathbb{R}^p$ (for details see Krzyśko (2006) or Jajuga (1993)).

A discriminant rule L corresponds to a division of $\mathbb{R}^p$ into disjoint regions $R_1, \ldots, R_k$, such that $\bigcup R_j = \mathbb{R}^p$. The rule L is definied by:

$$\text{Allocate } \mathbf{x} \text{ to } C_j \text{ if } \mathbf{x} \in R_j, \text{ for } j = 1, \ldots, k. \quad (1)$$

An index $i \in \{1, 2, \ldots, k\} = \mathrm{Y}$ corresponding to the population C_i is entitled as a label. In this setting a discrimination issue brings to a prediction of the label $i \in \mathrm{Y}$ on base o an measurement $\mathbf{x}$.

A classification rule is a function:

$$L\colon \mathbb{R} \ni \mathbf{x} \to i \in \mathrm{Y} \quad (2)$$

The function assigns to the vector $\mathbf{x} \in \mathbb{R}$ the prediction of the label $L(\mathbf{x}) \in \mathrm{Y}$.

[*] Ph.D., Department of Statistics, Cracow University of Economics.

The situation where the p.d.f.s $f_j(\mathbf{x})$ are known exactly is the simplest to analyze theoretically, although it is the least realistic in practice. A variant of this situation occurs when the form of the p.d.f. for each population is known, but there are parameters which must be estimated. Usually the estimation than is based on a sample data matrix (a training sample) $\mathbf{Z}_{n\times p}$, whose rows are partitioned into k groups corresponding k considered populations:

$$Z = \begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_k \end{bmatrix}, \tag{3}$$

where $(n_j \times p)$ matrix $\mathbf{Z}_j$ corresponds sample of n_j observations from C_j.

Classical discriminant methods like a linear or quadratic discriminant function assume multivariate normality of the populations. In case of skewed populations these methods have not optimal properties. The methods assume also an existence of first and second order moments what is useless in case of multivariate Cauchy distribution. Mean vector or covariance matrix based methods are extremely sensible to outliers. In this paper we propose a classification rule based on data depth concept that have less disadvantages than the classical classification methods.

II. ROBOUSTNESS OF A DISCRIMINATION RULE

A concept of breakdown points was introduced by Hodges and Hampel and still plays an important though at times a controversial role in robust statistics. It has proved most successful in the context of location, scale and regression problems. Below we propose an adaptation of the definition of the breakdown point adequate for classification issues.

Definition (compare with Krzyśko M. (2006)): Consider k p-dimensional populations $C_1, \ldots, C_k$, $k \geq 2$ and a fixed training sample **z** representing the populations. **An actual prediction error** of discriminative rule L is equal

$$Err(L) = P\{L(\mathbf{X}) \neq i \mid \mathbf{X} \in C_i, i = 1, \ldots, k\}, \tag{4}$$

where **X** denotes an observation independent on training sample.

Proposition 1: Consider k p-dimensional populations $C_1, \ldots, C_k$, $k \geq 2$ and a training sample **Z** representing these populations. **A breakdown point of the training sample Z of a classification rule** L **in** j **class** C_j is defined as

$$BP_j(L, C_j^m) = \inf_{C_j^m}\left(\frac{m}{n_j} : P\left\{L(\mathbf{X}) \neq j \mid \mathbf{X} \in C_j\right\} > 1/2\right), \tag{5}$$

where C_j^m denotes $(n_j \times p)$ sub matrix $\mathbf{Z}_j$ of the training sample **Z** corresponding a sample of n_j observation drawn from the population C_j where m rows (m observations) are replaced by arbitrary rows (outliers), **X** denotes an observation independent from the training sample **Z**.

An overall breakdown point of the training sample Z of the classification rule L is defined

$$BP(L, C_1, \ldots, C_k) = \min_j BP_j(L, C_j^m)\,. \tag{6}$$

To solve this classification problem we introduce classification rules that are based on a notion of a data depth. A depth function $d(\mathbf{x}, F)$ is a map from $\mathbb{R}^p$ into a subset of nonnegative real numbers, whose values provide a center-outward ordering of the points of $\mathbb{R}^p$ according to the probability distribution F. The highest value of $d(, F)$ corresponds to a center of the distribution (for details see Dyckerhoff R. (2004)).

Any depth function d provides a **depth classification rule**

$$L(\mathbf{z}) = \arg\max_j D(\mathbf{z} \mid C_j),\ j = 1, \ldots, k\,, \tag{7}$$

which assigns $\mathbf{z}$ to that class C_j in which $\mathbf{z}$ is deepest (see Hoberg & Mosler (2006)).

Let $C_j = \{\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jn_j}\}$ denotes a sample of n_j observations drawn from the population j, $j = 1, \ldots, k$.

Proposition 2: Consider a classification rule induced by a **symmetric projection depth function** (for details see Zuo (2003))

$$D_{PRO}(\mathbf{z} \mid C_j) = \left(1 + \sup_{\|\mathbf{u}\|=1} \frac{\left|\mathbf{u}'\mathbf{z} - med(\mathbf{u}'C_j)\right|}{MAD(\mathbf{u}'C_j)}\right)^{-1}, \tag{8}$$

where $\mathbf{u}'C_j = \left\{ \mathbf{u}'\mathbf{x}_{j1}, ..., \mathbf{u}'\mathbf{x}_{jn_j} \right\}$, $MAD(Y) = med\left\{ \left| Y - med(Y) \right| \right\}$.

The projection depth function is among others affine invariant and quasi convex. An induced location and scatter estimators have high finite sample replacement breakdown points and good properties in terms of Hampel's influence function and Huber's maximum bias (for details see Zuo (2003)).

A set

$$D^{\alpha}_{PRO}(\mathbf{x}^n) = \{\mathbf{z} \in D_{PRO}(\mathbf{z} \mid \mathbf{x}^n) \geq \alpha\}, \tag{9}$$

where $x^n = \{x_1, ..., x_n\} \subset \mathbf{R}^p$ denotes a sample, $0 \leq \alpha \leq 1$, is called an α **projection central region.**

The projection central regions constitute affine equivariant, nested and convex family of sets (for details see Dyckerhoff R. (2004)).

III. STATISTICAL FEATURES OF THE PROPOSED CLASSIFICATION RULE

Statistical properties of the proposed projection depth classification rule in comparison to linear and quadratic discriminant function was investigated using simulations and well known empirical data sets.

A. Table 1 shows a performance of the proposed classification rule on the well known data set consisting of measurements on three types of iris considered with respect to sepal length, sepal width, petal length and petal width (Fisher 1936). We considered 25:25:25 and 40:40:40 training sample sizes, on base of them all observations belonging to data set was classified. The results shows that in this case proposed procedure exhibits comparable properties to the classical methods.

Table 1. The results of the classification using proposed depth based procedure

	3x25 observations in the training sample			
	Linear discriminant function	Quadratic discriminant function	Projection depth based	Tukey depth based
Actual prediction error	2.6%	4%	3.3%	66%
	3x40 observations in the training sample			
Actual prediction error	3.8%	2.6%	2.6%	66%

B. We simulated 100 two dimensional data sets of sizes 3000 from the eqal size contamination of Marshall – Olkin distribution (1000) , isotropic normal distributions (1000) and skewed T Student with two degree of freedom distribution (1000). We draw training samples of sizes 100x100x100 from the data sets. On base of the training samples data sets were classified using the proposed procedure. Table 2 shows a performance of the proposed classification rule. The results shows that in this case proposed procedure exhibits much better properties to the classical methods.

Table 2. The results of the classification using proposed depth based procedure

	Classification rule			
	Linear discriminant function	Quadratic discriminant function	Projection depth based	Tukey depth based
Actual prediction error	12.6%	12.6%	0.3%	68%

Source: Our own calculations.

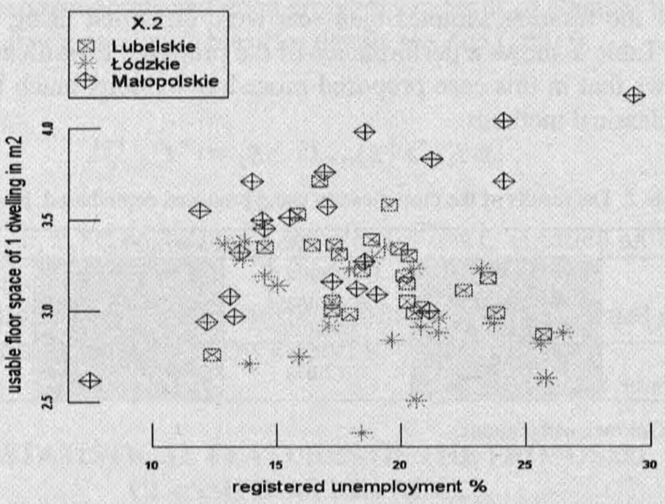
C. In order to estimate the overall breakdown point of the training sample we simulated data sets consisted of 3000 observations generated by equal contribution contamination of three skewed T Student distribution with different location and shape parameters. Next we replaced 0%,1%,...,10% of observations in 3x100 training sample drawn from simulated earlier data sets by outlying observations. We calculated the actual error of classification after that replacement. Table 3 shows the results for the fraction of outliers in the training sample varying from 0% to 10%. The results shows very good properties of the proposition in terms of robustness to outliers.

Table 3. The results of the overall breakdown point of the proposed depth based procedure estimation

Fraction of outliers in the training sample	0%	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
BP	0.004	0.008	0.017	0.023	0.03	0.043	0.04	0.05	0.05	0.07	0.07

Source: Our own calculations

D. Table 4 shows a performance of the proposed classification rule applied to an empirical data set consisted of 69 polish districts (powiats) from lubelskie province (24), łódzkie province (23) and małopolskie province (22) considered with respect to usable floor space of 1 dwelling in m^2 and net unemployment rate (figure 1). The results shows pour province classification properties of the proposition.



Pic. 1: Usable floor space of 1 dwelling in m^2 vs. net unemployment rate in 69 polish districts (powiats)

Source: Our own calculations, data GUS.

Table 4. The results of the classification using proposed depth based procedure

	Classification rule	
	Projection depth based	Tukey depth based
Actual prediction error	44.9%	68.1%

Source: Our own calculations

IV. ROBUST CLUSTERING PROCEDURE PROPOSITION

For an indication of further studies of the projection depth based classification procedures issues consider a following proposition of a robust clustering.

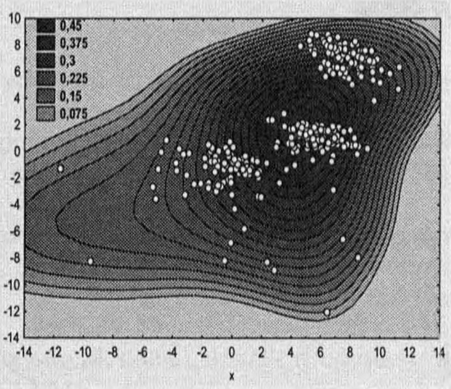
Let $C_0 = \{x_1, ..., x_n\}$ be measurements of n p-dimensional observations. Our aim is to group these objects into k-homogonous classes where k is also unknown. In other words our aim is to find an optimal partition C_0 into k homogonous disjoint subsets $C_1,...,C_k$, $k \geq 2$, $C_i \cap C_j = \varnothing$, $i \neq j$, $\bigcup C_i = C_0$.

Proposition 3: Let $\tilde{C}_1,...,\tilde{C}_k$, $k \geq 2$, $\tilde{C}_i \cap \tilde{C}_j = \varnothing$, $i \neq j$, $\bigcup \tilde{C}_i = C_0$ be certain possible partition of data set C_0. We call a partition $\tilde{C}_1,...,\tilde{C}_k$ better than a trivial partition C_0 and $\varnothing$ if

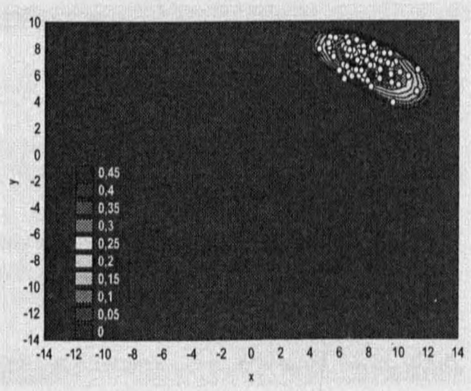
$$vol(D^{\alpha}_{PRO}(C_0)) > \sum_{i=1}^{k} vol(D^{\alpha}_{PRO}(\tilde{C}_i))\text{, for a fixed } \alpha \in (0,1), \qquad (10)$$

where $vol(D^{\alpha}_{PRO}(C))$ denotes a volume of α projection central region.

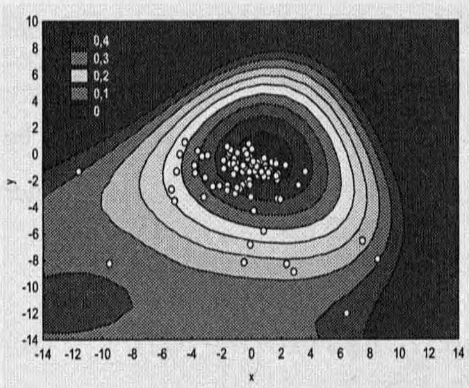
Pictures 1 – 4 shows an idea of clustering using the proposition. Picture 5 and clusters volumes calculations shows the problem of correct indication of $\alpha \in (0,1)$ parameter.



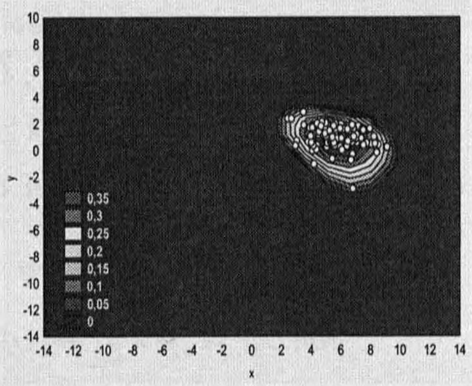
Pic 2. Depth based clustering procedure – clusters treated jointly – trivial partition



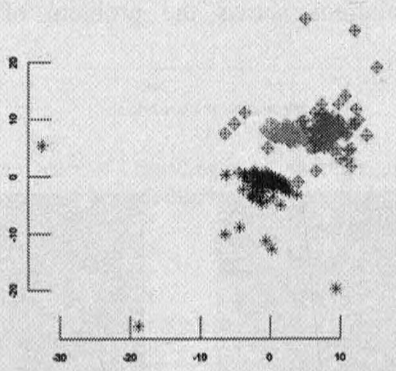
Pic 3. Depth based clustering procedure – depth in the first cluster



Pic 4: Depth based clustering procedure – depth in the second cluster



Pic 5: Depth based clustering procedure – depth in the third cluster



Pic 6. An illustration of a parametr choice in the depth based clustering procedure

cluster C_0 cluster $\tilde{C}_1$

$vol(D^{0.1}(C_0)) = 217.45cm^2$; $vol(D^{0.1}(C_1)) = 44.75cm^2$

$vol(D^{0.2}(C_0)) = 217.45cm^2$; $vol(D^{0.2}(C_1)) = 20.73cm^2$

cluster $\tilde{C}_2$ cluster $\tilde{C}_3$

$vol(D^{0.1}(C_2)) = 175.81cm^2$; $vol(D^{0.1}(C_3)) = 27.58cm^2$

$vol(D^{0.2}(C_2)) = 73.79cm^2$; $vol(D^{0.2}(C_3)) = 17.84cm^2$.

V. CONCLUSIONS

We presented projection depth based method for classification and the idea of projection depth clustering procedure. The simulation studies showed that classification rule proposition have good statistical properties in a context of the robustness. The proposition seems to be a competitive classifier to well known classifiers and others depth induced classification rules.

The proposed clustering procedure performed well on the simulated data sets. The clustering results analysis showed that our proposition was robust to a moderate fraction of outliers and generated clusters that could be used to predict sample labels better than k- means algorithm.

We are currently working on a further development of the proposed methods i.e. among others for a simplification of the computational aspects of the procedures (we focus our attention on the properties of a projection pursuit approach proposed by Dyckerhoff (2004)) and obtaining an idea about the number of natural clusters that are really present in the data sets (we study the possibilities of replacing a Schwarz information criterion by maximum depth criterion in a mixture based clustering modeling).

REFERENCES

Dyckerhoff R. (2004), *Data Depths Satisfying the Projection Property*, Allgemeines Statistisches Archiv, 88, p. 163–190.

Jajuga K. (1993), *Statystyczna analiza wielowymiarowa*, PWN, Warszawa.

Hoberg R., Mosler K. (2003), *Classification based on data depth*, Bulletin of the ISI 54th Session.

Hoberg R., Mosler K. (2006), Data analysis and classification with the zonoid depth, [in:] R. Liu, R. Serfling, D. Souvaine, eds., *Data Depth: Robust Multivariate Analysis*, Computational Geometry and Applications, American Mathematical Society, 2006, 49–59.

Krzyśko M. (2006), *Modele Klasyfikacyjne*, referat plenarny na konferencji MSA 2006, Łódź.

Rousseeuw P. J., Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, Willey, New York.

Zuo Y. (2003), *Projection Based Depth Functions and Associated Medians*, The Annals of Statistics, 31(5): 1460–1490, 2003.

Daniel Kosiorowski

ODPORNOŚĆ METOD KLASYFIKACYJNYCH WYKORZYSTUJĄCYCH FUNKCJE GŁĘBI

W referacie proponujemy kilka reguł klasyfikacyjnych wykorzystujących funkcje głębi. Badamy ich właściwości m. in. na różnych zbiorach danych generowanych przez wielowymiarowe skośne rozkłady, rozkłady o tłustych ogonach i mieszaniny takich rozkładów. Dyskutujemy także regułę pozwalającą na wybór właściwej liczby klas $A_1,\ldots,A_k$ stanowiących podział zbioru danych A.