Austrian Journal of Statistics
April 2017, Volume 46, 117–128.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.677

Depth-based Classification for Multivariate Data

Ondřej Vencálek
Palacký University in Olomouc

Abstract

Concept of data depth provides one possible approach to the analysis of multivariate data. Among other it can be also used for classification purposes. The present paper is an overview of the research in the field of depth-based classification for multivariate data. It provides a short summary of current state of knowledge in the field of depth-based classification followed by detailed discussion of four main directions in the depth-based classification, namely semiparametric depth-based classifiers, maximal depth classifier, (maximal depth) classifiers which use local depth functions and finally advanced depth-based classifiers. We do not restrict our attention only on proposed classifiers. The paper rather aims to overview the ideas connected with depth-based classification and problems that were discussed in this context.

Keywords: data depth, classification, Bayes optimal, overview.

1. Introduction

Depth function is basically any function which provides ordering (or quasi-ordering) of points in multidimensional space $\mathbb{R}^d$ with respect to some probability measure P defined on this space. Existence of ordering enables generalization of quantiles (median, in special case) and related nonparametric techniques proposed for univariate variables. Thus the notion of depth creates one possible basis of nonparametric multivariate data analysis.

Data depth has been also applied in classification. We use the term classification, where sometimes term discriminant analysis or supervised learning is used. The aim of classification is to create a rule for allocation of new observations into one of two (or more) groups. Formally, we consider two unknown absolutely continuous probability distributions P_1 and P_2 on $\mathbb{R}^d$. Independent random samples from these distributions are available. Together they constitute so called training set. Empirical distributions based on the training set are denoted $\widehat{P}_1$ and $\widehat{P}_2$. A classifier (rule for classification) is thus a function $c : \mathbb{R}^d \to \{1, 2\}$. The theory can be generalized for more than two groups.

Possibility to use data depth for classification was firstly mentioned by Liu already in 1990, Liu (1990). She suggested that the new observation "should be assigned to the population whose training sample leads to a smaller relative rank for it." However, after this first reference it lasted more than ten years until the depth-based classifiers started to be studied systematically. Thus we can say that the depth-based classification has been developed since 2000.

The aim of the present paper is to summarize main ideas how the depth can be used in classification. The paper renders readers interested in classification insight to the current state of knowledge in the field of depth-based classification. Also it provides a comprehensive overview of the research made in the area of depth-based classification in the last 15 years.

2. Short guide on depth-based classification

The reader not familiar with the depth-based classification might ask in which situation it was useful to use this concept. In order to facilitate the basic orientation in this area we shortly summarize current state of knowledge in the field of depth-based classification. We do it even before dealing with particular methods of depth-based classification since it might be useful for the reader to know what to expect from the depth-based classification as soon as possible.

- The depth-based classifiers are applicable when one want to avoid strict parametric assumptions (like normality) on the considered distributions. They can utilize possible global properties of the considered distributions (like their symmetry), but can be applied even if there are no such properties, e.g. for non-symmetric distributions. They work well for unimodal distributions since the depth is global property which characterizes location of a point w.r.t. the whole distribution. If the considered distributions might be multimodal or could have nonconvex levelsets of density a local depth should be used to overcome this problem.
- Simple advice how to decide weather it is reasonable to use depth-based classification is to answer the question weather it is reasonable to classify by median and other quantiles. The median which is the point with highest depth might lie in the area where the density is low. In such a case depth-based classifiers would have problems. Only few of them are able to overcome these problems.
- Depth-based classifiers were primarily constructed for continuous distributions, little attention was paid for categorical "explanatory" variables.
- One should be aware that the computation of depth is not easy task and for most of the depth functions might be very slow in dimensions higher than five when the number of points in training set is high. The depth-based procedures are advisable in dimensions from 2 to 5, they are not advisable in dimensions higher than 20. However, the latest depth-based classifiers, as the one proposed by Dutta and Ghosh (2015), were shown to perform well even in dimension 100. For really high-dimensional data the depth-based classifiers usually need to be preceded by reduction of dimensionality.
- When more than two classes are considered, the depth-based classifiers usually rely on majority voting principle.
- The main advantage of the depth-based classifiers is their affine invariance. Most of the depth functions are affine invariant and thus the classification procedures do not change e.g. with the change of units (scales). The depth-based classifiers also usually have good robust properties.
- The simplest classifiers, so called maximal depth classifiers, which will be described in section 4.2 are not satisfactory in most practical situations. More complicated classifiers described in section 4.4 need to be employed. Probably the most universal depth-based classifier is that by Paindaveine and Van Bever (2015) (see end of the section 4.4) since it is shown to be nonparametrically consistent under very mild conditions.
- For practitioners who need reliable implemented classifiers an R-package ddalpha is advisable, see Pokotylo, Mozharovskyi, and Dyckerhoff (2016). In this package the DD-plot classifier by Li, Cuesta-Albertos, and Liu (2012) and the DDα-classifier by Lange,

Mosler, and Mozharovskyi (2014b) are implemented. Both these classifiers (mentioned in section 4.4) belong to the best depth-based classifiers that are currently available.

- There are also depth-based methods for classification of functional data. They utilize so called functional data depth. The current paper deals with classification for multivariate data so the classification of functional data is not included.

3. Concept of data depth

There are several depth functions commonly used for classification – halfspace depth, projection depth, spatial depth, Mahalanobis depth, zonoid depth, and some others. Let us recall here the first four of the depth functions listed above:

- The *halfspace depth* of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P is defined as the minimum probability mass carried by any closed halfspace containing $\boldsymbol{x}$, that is

$$D(\boldsymbol{x}, P) = \inf\left\{\mathrm{P}(\mathbb{H}) : \mathbb{H} \text{ a closed halfspace in } \mathbb{R}^d : \boldsymbol{x} \in \mathbb{H}\right\}.$$

- The *projection depth* of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P is defined as

$$D(\boldsymbol{x}, P) = \frac{1}{1 + O(\boldsymbol{x}, P)}, \quad \text{where} \quad O(\boldsymbol{x}, P) = \sup_{\|\boldsymbol{u}\|=1} \frac{|\boldsymbol{u}'\boldsymbol{x} - \mu_{P_{\boldsymbol{u}}}|}{\sigma_{P_{\boldsymbol{u}}}},$$

  where $\mu_{P_{\boldsymbol{u}}}$ is some location and $\sigma_{P_{\boldsymbol{u}}}$ is some scale measure of distribution of random variable $\boldsymbol{u}'\boldsymbol{X}$ ($\boldsymbol{X} \sim P$), usually $\mu_{P_{\boldsymbol{u}}} = \mathrm{median}(\boldsymbol{u}'\boldsymbol{X})$ and $\sigma_{P_{\boldsymbol{u}}} = \mathrm{MAD}(\boldsymbol{u}'\boldsymbol{X})$, where MAD stands for median absolute deviation.

- The *Mahalanobis depth* of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ is defined as

$$D(\boldsymbol{x}, P) = \frac{1}{1 + O(\boldsymbol{x}, P)}, \quad \text{where} \quad O(\boldsymbol{x}, P) = (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}).$$

- The *spatial depth* (also called L_1*-depth*) of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P with variance matrix $\boldsymbol{\Sigma}$ is defined as

$$D(\boldsymbol{x}, P) = 1 - \mathrm{E}_P \left\| \frac{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{X})}{\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{X})\right\|} \right\|.$$

All the depth functions listed above have desirable properties like affine invariance, maximality at a point of symmetry (if the distribution is symmetric in some sense, e.g. angularly), monotonicity on rays from the point with the maximal depth – so called deepest point, which can be considered as multivariate analogy to median.

The very important difference among the considered depth functions consists in different behaviour of their empirical versions. While the empirical halfspace depth is equal to zero for any point which lies outside of the convex hull of the data, the empirical versions of the latter three depth functions are nonzero everywhere.

The depth of a point is a characteristic of the point specifying its centrality or outlyingness with respect to the considered distribution. Since the whole distribution is considered, the depth is said to be a "global" characteristic of the point. However, in recent years there have been attempts to "localize" depth. Later we will discuss importance of these attempts for classification purposes.

4. Depth-based classifiers

We can distinguished four main groups of depth-based classifiers for multidimensional data: semiparametric classifiers which use data depth, maximal depth classifier which use global depth functions, maximal depth classifiers which use local depth functions and "advanced" depth-based classifiers.

4.1. Semiparametric depth-based classifiers

- The first thorough study devoted to possible use of data depth in context of classification entitled "Measuring overlap in binary regression" was done by Christmann and Rousseeuw (2001). The authors considered classical logistic regression model in which the 0-1 response variable coded group membership in two classes classification problem. They realized that the regression depth can be used to measure amount of separation between the two groups. Enumeration of the regression depth of one particular 'fit' is accompanied by finding the hyperplane which minimizes number of misclassified points from the training set. Similar ideas can be found also in Christmann, Fischer, and Joachims (2002) and Christmann (2006).

- Ghosh and Chaudhuri extended substantially the above mentioned ideas, see Ghosh and Chaudhuri (2005a). They came with the following extensions: (1) They found connection of the linear classification with the halfspace depth ("the estimated linear projection is orthogonal to the hyperplane, which defines the halfspace depth of the origin w.r.t. the data cloud formed by the differences $\boldsymbol{x}_{1i} - \boldsymbol{x}_{2j}$ in the d-dimensional space", where $\boldsymbol{x}_{1i}$ are observations from one group and $\boldsymbol{x}_{2j}$ from the other). (2) They suggested use of weighed regression depth in linear classification based on regression depth to deal with the situation in which prior probabilities are not proportional to their training sample sizes. (3) They generalized the classifiers for nonlinear separating surfaces. To construct such surfaces they projected the original d-dimensional observations $\boldsymbol{x}_i$ into a higher-dimensional space of features $\boldsymbol{z}_i = (f_1(\boldsymbol{x}_i), \ldots, f_h(\boldsymbol{x}_i))$, where $f_1(\cdot), \ldots, f_h(\cdot)$ are some given functions, e.g. powers, and performed linear classification on that h-dimensional space. In this way they could obtained e.g. quadratic classification. Similar idea, but in slightly different context can be found in Lange *et al.* (2014b). (4) They discussed the problem of more than two groups. They proposed majority voting or pairwise coupling.

4.2. Maximal depth classifier

- The simplest depth-based classifier is so called maximal depth classifier. The points close to the centre (multivariate median) of some distribution have high depth with respect to this distribution and it seems to be natural to classify them to this distribution. The idea of the maximal depth classifier – to classify a new observation to the group where its depth is maximal – is thus in accordance to common sense:

$$c(\boldsymbol{x}) = \arg\max_{i=1,2} D(\boldsymbol{x}; \widehat{P}_i) \tag{1}$$

  Different authors advocated use of different depth functions in this classifier. As far as we know this classifier was first studied by Jörnsten (2004). She used spatial depth, similarly as Hartikainen and Oja (2006). Jörnsten also came with the idea of relative depth which measures uncertainty of classification of a given point. The relative depth is defined as difference between maximal and second highest depth, i.e. $\max_i D(\boldsymbol{x}; \widehat{P}_i) - \max_{i \neq c(\boldsymbol{x})} D(\boldsymbol{x}; \widehat{P}_i)$. If the difference is high we are pretty sure in classification. Small relative depth indicates possible problems in the class assignment. Jörnsten suggested deletion of points with small relative depth from the training sample

which might lead to improvement in classification. The classifier proposed by Jörnsten was used in comparative study focused on classifiers for high-dimensional data presented in Hall, Titterington, and Xue (2009). However, this study highlighted rather componentwise median-based classifier and its truncated form.

Mosler and Hoberg were first who pointed out so called "outsider" problem in Mosler and Hoberg (2006). The problem consists in zero empirical depth of points that are outside of the convex hull of points in the training set when using some depth functions like zonoid depth or halfspace depth. To overcome the outsider problem they suggested to combine zonoid depth with Mahalanobis depth which does not suffer from this problem.

Use of projection depth in the maximal depth classifier was advocated by Kosiorowski (2008) who emphasized good robust properties of such a classifier. Also the classifier proposed by Hubert and Van der Veeken (2010b) can be understand as the maximal depth classifier using projection depth where the outlyingness adjusted for skewness of the distribution is used instead of commonly used outlyingness. Similarly as Jörnsten they consider removal of the possible problematic points from the training set before the construction of the final classifier. Their contribution can be viewed also in (rather exceptional) discussion of depth-based classification for high-dimensional data. They advocated use of robust SIMCA (Soft Independent Modelling by Class Analogy) method which lies in application of (robust) PCA method in each group. An extended projection depth which takes into account possible difference in dispersion of considered distributions was suggested by Cui, Lin, and Yang (2008).

An interesting technical application of maximal depth classifier for real time sensor node tracking and location was recently proposed in Kumar, Kumar, Kumar, and Hegde (2015). Specificity of this classifier lies in the fact that not one point, but the group of multidimensional points is classified at once.

- Although many authors suggesting maximal depth classifier refered to the original paper by Liu (1990), the idea presented in that paper was to classify rather according to the relative rank (based on depth), not the depth itself (see section 1). This idea was rediscovered much later by Billor, Abebe, Turkmen, and Nudurupati (2008). Although they call their classifier as "depth transvariation classifier", it can be called more directly maximal rank classifier since it has the following form:

$$c(\boldsymbol{x}) = \arg\max_{i=1,2} \operatorname{rank}(\boldsymbol{x}; \widehat{P}_i), \tag{2}$$

where $\operatorname{rank}(\boldsymbol{x}; \widehat{P}_i)$ denotes percentage of points from the training set of the i-th group which have smaller depth w.r.t. $\widehat{P}_i$. The same idea was also considered by Hubert and Van der Veeken (2010a).

The maximal depth classifiers mentioned in this section are already outdated and were replaced by maximal depth classifiers that use some local depth described in section 4.3 or by more advanced classifiers described in section 4.4. Insufficiency of the maximal depth classifiers was revealed already by Ghosh and Chaudhuri (2005b). They proved that the maximal depth classifier attains minimal possible probability of misclassification (known as Bayes risk) only in very special cases – they showed optimality when the considered distributions are elliptically symmetric with the density decreasing from the centre, differing only in location and having equal prior probabilities. The optimality is lost even if only one of these assumptions is not fulfilled. The following simple example illustrates these problems. Let us consider two one-dimensional normal distributions with the same mean but different variances. Then all points different from the mean will be classified to the distribution with smaller variance. This is due to the affine invariance of the depth.

4.3. Maximal depth classifiers with local depth

The depth of a given point characterizes its location w.r.t. the whole distribution. Thus the classifiers which use any "global" depth function perform well only if the considered distributions have some global properties like symmetry or unimodality. To obtain good performance also in more general settings, for example in the case of multimodality or nonconvexity of levelsets of density, use of some local depth started to be promoted recently. A new problem that emerged and need to be handle is choice of localization level.

The first classifiers which employed local depth appeared only in 2013. Hlubinka and Vencalek (2013) used weighted halfspace depth. Paindaveine and Van Bever (2013) developed more complex approach which enables localization of any "global" depth function which can be subsequently used in the maximal depth classifier. The proposed local depth is defined as global depth conditional on some neighborhood of the point of interest. The neighborhood itself is defined in terms of data depth. It is worthwhile to recall the main ideas leading to the localization of depth here:

Let $P^{\boldsymbol{X}}$ be a probability distribution of a d-dimensional random variable $\boldsymbol{X}$, let $D(\cdot, P^{\boldsymbol{X}})$ be any depth function and let $\beta \in (0, 1]$. *Depth regions* of a distribution $P^{\boldsymbol{X}}$ are sets of the following form: $\left\{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}, P^{\boldsymbol{X}}) \geq \alpha\right\}$. The *symmetrized version of a distribution* $P^{\boldsymbol{X}}$ with the centre in a given point $\boldsymbol{x} \in \mathbb{R}^d$ is defined as a mixture $P_{\boldsymbol{x}} = \frac{1}{2}P^{\boldsymbol{X}} + \frac{1}{2}P^{2\boldsymbol{x}-\boldsymbol{X}}$. *The probability β neighborhood of a point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $P^{\boldsymbol{X}}$* is defined as the smallest depth region of $P_{\boldsymbol{x}}$ with $P_{\boldsymbol{x}}$ probability larger than or equal to β. It is denoted by $R^{\beta}(P_{\boldsymbol{x}})$. The *$\beta$-local depth of a point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $P^{\boldsymbol{X}}$* is defined as $D(\boldsymbol{x}, P^{\beta}_{\boldsymbol{x}})$, where $P^{\beta}_{\boldsymbol{x}}$ is conditional distribution of $P^{\boldsymbol{X}}$, conditional on depth neighborhood of $\boldsymbol{x}$ $R^{\beta}(P_{\boldsymbol{x}})$.

An interesting economical application of maximal depth classifier which uses local L^p depth for so called algorithmic trading was presented by Kosiorowski, Bocian, and Bujak (2014). The classes considered in the paper characterize different states of market.

4.4. Advanced depth-based classifiers

The paper by Ghosh and Chaudhuri (2005b) uncovered insufficiency of the maximal depth classifier and started the search for depth-based classifiers which would be applicable in a broad class of distributional settings. Typical depth-based classifier can be described as a two-steps procedure:

1. The first step consists in computation of depths of the new observation $\boldsymbol{x}$ with respect to both parts of the training set. Each point is characterized by a pair of depths, these pairs lies in so called DD-space (depth-versus-depth space). Typically the DD-space is subset of $[0, 1] \times [0, 1] \subset \mathbb{R}^2$ and thus the first step can be usually (but not necessarily) considered as reduction of dimensionality – from $\mathbb{R}^d$ to the compact subset of $\mathbb{R}^2$. This step is connected with the question "Which depth function should be used?"

2. The second step consists in application of some classification procedure in the DD-space. This step is connected with the question "Which classification procedure should be applied in the DD-space?" This "new" question is in the center of current research in the field of depth-based classification.

The scheme of typical depth-based classification procedure is shown in Figure 1.

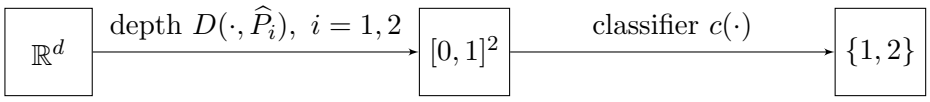

Figure 1: Scheme of typical advanced depth-based classifier.

The difference among the classifiers proposed in literature consists mainly in different answers to the two questions connected with the two steps – different depth functions can be applied in

the first step and different classification procedures can be applied in the second step. Depth function used in the first step can be either global or local. Also procedures that are applied in the DD-space can be also either "global" or "local" in nature. As the global we denote the procedures that take into account all points of the training set when constructing the classifier while the local procedures take into account only points of the training set close to the point which is classified.

The list of advanced depth-based classifiers follows:

- The first classifier which overcame insufficiency of the maximal depth classifier was suggested already by Ghosh and Chaudhuri (2005b). They discovered that in the case of elliptically symmetric distribution the classifier which minimizes probability of misclassification can be expressed as

$$c(\boldsymbol{x}) = \arg\max_{i=1,2} \pi_i \theta_i(D(\boldsymbol{x}, \widehat{P}_i)),$$

  where π_i are prior probabilities and $\theta_i, i = 1, 2$ are some unknown real functions. This holds due to the correspondence between depth and density in case of elliptically symmetric distributions. They found explicit formula for the $\theta_i(D(\boldsymbol{x}, \widehat{P}_i))$, when the halspace depth is used. Later Dutta and Ghosh (2012) found similar formula also for the projection depth. The classifiers which utilize these relations have the following form:

$$\begin{aligned} c(\boldsymbol{x}) &= \arg\max_i k_i \rho_i(\gamma_i(HD(\boldsymbol{x}, \widehat{P}_i)))/\gamma_i(HD(\boldsymbol{x}, \widehat{P}_i))^{d-1}, \\ c(\boldsymbol{x}) &= \arg\max_i k_i^* \rho_i^*(PD(\boldsymbol{x}, \widehat{P}_i)) \cdot PD(\boldsymbol{x}, \widehat{P}_i)^{d-3}/(1 - PD(\boldsymbol{x}, \widehat{P}_i))^{d-1}, \end{aligned}$$

  where the first classifier uses halfspace depth HD while the second classifier uses projection depth PD, k_i (k_i^* respectively) are unknown constants estimated by minimizing misclassification rate, $\gamma_i(HD(\boldsymbol{x}, \widehat{P}_i))$ denotes Mahalanobis distance whose relation to the halfspace depth is known, and finally ρ_i (ρ_i^* respectively) are unknown functions that need to be estimated by kernel density estimation technique. Here it is useful to note that the density that need to be estimated is always one-dimensional.

- The previously mentioned classifiers are optimal if the considered distributions are l_2-symmetric (after standardization). However, their authors have shown that for any l_p-symmetric distribution with $p \neq 2$ the density cannot be a function of halfspace depth. Dutta and Ghosh (2016) suggested to use L_p depth ($D(\boldsymbol{x}, P) = 1/(1 + \left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})\right\|_p)$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are location and scale parameters of the distribution P) to overcome this problem and obtain classifier optimal for a broader class of l_p-symmetric distributions. They used relationship between depth and density in this case to estimate the density from the depth by one-dimensional kernel density estimation. The choice of p had to be discussed. The authors proposed restricted maximal likelihood estimate and discussed further how the restriction should be made.

- Another interesting classifier was proposed by Dutta and Ghosh (2015). They first transform the original data to the DD-space using spatial or localized spatial depth. Subsequently they use these depths as explanatory variables in multinomial additive logistic regression model (a spacial case of generalized additive models, called GAMs) where the response variable indicates group membership (class labels). Degree of localization determined by a single parameter h of locality is object of interest. The authors suggest "multiscale approach" in which the final classifier is based on weighted average of posterior probabilities estimated with different values of the parameter h. The weights are based on estimated misclassification rates. As mentioned already in section 2, this classifier was tested for high-dimensional data and demonstrated its good properties in this settings.

- The *DD-plot classifier* proposed by Li *et al.* (2012) works also in two steps. The first is transformation of the data into DD-space (depth versus depth). Visualization of the DD-space is by means of so called DD-plot. This step is followed by separating the transformed data points by a curve from a given family, for example by a polynom. The separation is done in a way which minimizes number of errors when classifying points from the training set. The classifier thus has the following form:

$$c(\boldsymbol{x}) = \arg\max_i r(D(\boldsymbol{x}, \widehat{P}_i)),$$

  where r is a function from a given class of functions (e.g. polynoms). In the simplest case the points might be separated by the straight line going through origin (which represents points with zero empirical depth to each group, so called outsiders). The only parameter that need to be estimated here (by minimizing average error rate on training set) is the slope of the line. In a special case we obtain maximal depth classifier which corresponds to the 45 degree line.

  The idea to separate classes by straight line going from the origin in the DD-space was proposed already by Jin and Cui (2010). Unfortunately, their paper remains unnoticed so far. They suggested (and subsequently used) new notion of depth, which is defined as $D_c(\boldsymbol{x}) = D(c\boldsymbol{x} + (1-c)\boldsymbol{\mu}, P)$, where D is some well known depth function and $c > 0$ is a tuning parameter. Apart from the slope of the line this parameter need to be estimated. For a given tuning parameter c the slope b is estimated to make probability of misclassification rate in the first group smaller than a given small $\alpha \in (0,1)$, i.e. $P_1(D(\boldsymbol{x}, \widehat{P}_1) > bD(\boldsymbol{x}, \widehat{P}_2)) \geq 1 - \alpha$. The parameter c is then tuned to make the misclassification rate in the second group as small as possible. Different notions of depth can be used to further minimize this misclassification rate.

- The *DD-alpha* procedure proposed by Lange *et al.* (2014b) belongs to the best currently available depth-based classifiers. Instead of the pair $[D(\boldsymbol{x}, \widehat{P}_1), D(\boldsymbol{x}, \widehat{P}_2)]$, it works with a higher-dimensional vector of "features". Such a vector might be like this:

$$\boldsymbol{z} := [D(\boldsymbol{x}, \widehat{P}_1), D(\boldsymbol{x}, \widehat{P}_2), D(\boldsymbol{x}, \widehat{P}_1) \cdot D(\boldsymbol{x}, \widehat{P}_2), D(\boldsymbol{x}, \widehat{P}_1)^2, D(\boldsymbol{x}, \widehat{P}_2)^2].$$

  Note that the the particular features always have the form of product $D(\boldsymbol{x}, \widehat{P}_1)^{k_1} \cdot D(\boldsymbol{x}, \widehat{P}_2)^{k_2}$, where in the pravious example $k_1 + k_2 \leq 2$, but in general higher powers can be used as well. Linear separation (by a hyperplane) in the feature space leads in general to nonlinear separation in the DD-space. Lange et al. proposed a heuristic for finding proper parameters which specify separating hyperplane given by the equation $aD(\boldsymbol{x}, \widehat{P}_1) + bD(\boldsymbol{x}, \widehat{P}_2) + cD(\boldsymbol{x}, \widehat{P}_1)D(\boldsymbol{x}, \widehat{P}_2) + dD(\boldsymbol{x}, \widehat{P}_1)^2 + eD(\boldsymbol{x}, \widehat{P}_2)^2 = 0$. The procedure was successfully tested on many real datasets leading usually to low misclassification rates, see Mozharovskyi, Mosler, and Lange (2015). The procedure which is very fast and robust was implemented in the R-package `ddalpha`, see Pokotylo *et al.* (2016). Several depth functions are implemented in this package. The choice of proper depth function was discussed in Lange, Mosler, and Mozharovskyi (2014a). The research on DD-alpha procedure is nicely summarized in Mozharovskyi (2015).

- The *k-depth-nearest neighbour classifier* highlighted by Vencalek (2013) is quite simple – it uses the well known k-nearest neighbour procedure in the DD-space. The question is which metric should be used to measure distances between distinct points.

- A classifier which uses a specific local depth was suggested by Pokotylo and Mosler (2016). Instead of term "DD-plot" used by Li *et al.* (2012) they use term "pot-pot plot", where pot-pot is a shortcut for potential versus potential. The potential of a class in a given point is defined as a kernel density estimate in this point multiplied by the class's prior probability. Proper choice of kernel can make the potential be affine invariant and thus it can be viewed as a local depth. Any of previously mentioned classifiers can be applied in the pot-pot plot.

- The transformation used in the first step of advanced depth-based classifier does not have to be directly transformation to the DD-space. Hubert, Rousseeuw, and Segaert (2015) suggested transformation to the distance-distance space. The considered distance is so called bagdistance which is based on (halfspace) depth. The bagdistance of a point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. distribution P is given by the ratio of the Euclidean distance of $\boldsymbol{x}$ to the multivariate median $\boldsymbol{\theta}$ and the Euclidean distance of a point $c(\boldsymbol{x})$ to $\boldsymbol{\theta}$, where $c(\boldsymbol{x})$ is defined as the intersection of the boundary of so call "bag" and a ray from the $\boldsymbol{\theta}$ through $\boldsymbol{x}$. The bag is the smallest depth region (for definition see section 4.3) with at least 50% probability mass. The ratio is not defined in $\boldsymbol{\theta}$, where the distance is defined additionally to be zero. In the distance-distance space they proposed to use k-nearest neighbour method, however any other classifier mentioned in this section can be used as well.

- There are two depth-based advanced classifiers that does not follow the scheme presented in Figure 1. Although they are different they both can be called k-depth nearest neighbours (k-depthNN) method.

  The first one was proposed by Vencalek (2011). The classifier was based on assumption that there exists a function which relates depth and density function. This assumption holds true for elliptically symmetric distributions. In other cases localization of depth might be used to bring the depth closer to density. Vencalek defined distributional neighbourhood of a given point $\boldsymbol{x} \in \mathbb{R}^d$ as a set of points whose depth w.r.t. a given distribution P does not differ from $D(\boldsymbol{x}, P)$ of more that a given $\epsilon > 0$. In analogy to classical kNN he considered points in neigbouhoods of $\boldsymbol{x}$ that contain a fixed number (k) of points. Since there are two distributions, there are also two such distributional neighbourhoods. Vencalek suggested to classify a new observation to the group with the smallest (in the sense of Lebesgue measure) distributional neighbourhood. The main practical problem of the procedure is computation (estimation) of Lebesgue measures of the distributional neighbourhoods.

  Maybe the most promising depth-based classifier is that by Paindaveine and Van Bever (2015). They used the idea of symmetrization: any given point $\boldsymbol{x} \in \mathbb{R}^d$ is in the centre of the equal mixture of original distribution $P^{\boldsymbol{X}}$ and its reflection $P^{2\boldsymbol{x}-\boldsymbol{X}}$ (see also section 4.3) and thus it is the deepest point w.r.t. this mixture. The points from the original training set can be ordered according to their depth w.r.t. this symmetrized distribution. Then k points with the highest depth form closest distributional neighbours of the point $\boldsymbol{x}$. The point is assigned to the group with the highest number of representatives among the k nearest neighbours. The procedure was shown to be "nonparametrically consistent" (which is just a little bit weaker property than universal consistency). The main practical disadvantage might be viewed in large number of computation needed to classify a single point.

5. Conclusion

The data depth provides basis for nonparametric inference on multidimensional data. Possibility of its use in classification has been investigated for more than 15 years. Although one can expect broad applicability of the nonparametric depth-based classifiers the optimality of many proposed classifiers can be guaranteed only under some restrictive assumptions. Global depth functions and global classification techniques applied on the DD-space lead to good results only if the considered distributions have some global properties like unimodality. In more general settings localization is needed – one can use local depth functions or local classifiers used in the DD-space. Simple classifiers like maximal depth classifier have been already overcome. The DD-plot classifier by Li *et al.* (2012), DD-alpha classifier by Lange *et al.* (2014b) (both implemented in the R-package ddalpha) and classifiers based on symmetrization proposed in Paindaveine and Van Bever (2015) and Paindaveine and Van Bever

(2013) belong to the currently top depth-based classifiers. Interesting new ways how to use depth in the context of classification like the one proposed in Gilad-Bachrach and Burges (2013) continue to appear.

Acknowledgement

The research was supported by the grant of Czech Science Foundation GA15-06991S.

References

Billor N, Abebe A, Turkmen A, Nudurupati SV (2008). "Classification Based on Depth Transvariations." *Journal of classification*, **25**(2), 249–260.

Christmann A (2006). "Regression Depth and Support Vector Machine." *DIMACS series in discrete mathematics and theoretical computer science*, **72**, 71–86.

Christmann A, Fischer P, Joachims T (2002). "Comparison between Various Regression Depth Methods and the Support Vector Machine to Approximate the Minimum Number of Missclassifications." *Computational Statistics*, **17**(2), 273–287.

Christmann A, Rousseeuw PJ (2001). "Measuring Overlap in Binary Regression." *Computational Statistics & Data Analysis*, **37**(1), 65–75.

Cui X, Lin L, Yang G (2008). "An Extended Projection Data Depth and Its Applications to Discrimination." *Communications in Statistics – Theory and Methods*, **37**(14), 2276–2290.

Dutta S, Ghosh AK (2012). "On Robust Classification Using Projection Depth." *Annals of the Institute of Statistical Mathematics*, **64**(3), 657–676.

Dutta S, Ghosh AK (2015). "Multi-scale Classification Using Localized Spatial Depth." *arXiv preprint arXiv:1504.03804*.

Dutta S, Ghosh AK (2016). "On Affine Invariant *L_p* Depth Classifiers based on an Adaptive Choice of *p*." *arXiv preprint arXiv:1611.05668*.

Ghosh AK, Chaudhuri P (2005a). "On Data Depth and Distribution-free Discriminant Analysis Using Separating Surfaces." *Bernoulli*, pp. 1–27.

Ghosh AK, Chaudhuri P (2005b). "On Maximum Depth and Related Classifiers." *Scandinavian Journal of Statistics*, **32**(2), 327–350.

Gilad-Bachrach R, Burges CJ (2013). "Classifier Selection Using the Predicate Depth." *Journal of Machine Learning Research*, **14**(1), 3591–3618.

Hall P, Titterington D, Xue JH (2009). "Median-Based Classifiers for High-Dimensional Data." *Journal of the American Statistical Association*, **104**(488), 1597–1608.

Hartikainen A, Oja H (2006). "On Some Parametric, Nonparametric and Semiparametric Discrimination Rules." *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **72**, 61–70.

Hlubinka D, Vencalek O (2013). "Depth-based Classification for Distributions with Nonconvex Support." *Journal of Probability and Statistics*, **2013**.

Hubert M, Rousseeuw PJ, Segaert P (2015). "Multivariate and Functional Classification Using Depth and Distance." *arXiv preprint arXiv:1504.01128*.

Hubert M, Van der Veeken S (2010a). "Fast and Robust Classifiers Adjusted for Skewness." In *Lechevallier, Y., Saporta, G. (Eds.), Proceedings of COMPSTAT 2010*, pp. 1135–1142. Physica-Verlag.

Hubert M, Van der Veeken S (2010b). "Robust Classification for Skewed Data." *Advances in Data Analysis and Classification*, **4**(4), 239–254.

Jin J, Cui H (2010). "Discriminant Analysis Based on Statistical Depth." *Journal of Systems Science and Complexity*, **23**(2), 362–371.

Jörnsten R (2004). "Clustering and Classification Based on the L_1 Data Depth." *Journal of Multivariate Analysis*, **90**(1), 67–89.

Kosiorowski D (2008). "Robust Classification and Clustering Based on the Projection Depth Function." In *Brito, P. (Eds.), Proceedings of COMPSTAT 2008*, volume II, pp. 209–216. Physica-Verlag.

Kosiorowski D, Bocian M, Bujak A (2014). "A Combination of Localdepth and svm Algorithms in Automatic Identification and Prediction of a Market State." In *Knowledge-Economy-Society: Contemporary Tools of Organizational Resources Management*, pp. 225–237. Cracow University of Economics.

Kumar S, Kumar A, Kumar A, Hegde RM (2015). "Hybrid Maximum Depth-kNN Method for Real Time Node Tracking Using Multi-sensor Data." In *2015 IEEE International Conference on Communications (ICC)*, pp. 6652–6657. IEEE.

Lange T, Mosler K, Mozharovskyi P (2014a). "DDα-classification of Asymmetric and Fat-tailed Data." In *Data Analysis, Machine Learning and Knowledge Discovery*, pp. 71–78. Springer.

Lange T, Mosler K, Mozharovskyi P (2014b). "Fast Nonparametric Classification Based on Data Depth." *Statistical Papers*, **55**(1), 49–69.

Li J, Cuesta-Albertos JA, Liu RY (2012). "DD-classifier: Nonparametric Classification Procedure Based on DD-plot." *Journal of the American Statistical Association*, **107**(498), 737–753.

Liu RY (1990). "On a Notion of Data Depth Based on Random Simplices." *The Annals of Statistics*, **18**(1), 405–414.

Mosler K, Hoberg R (2006). "Data Analysis and Classification with the Zonoid Depth." *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **72**, 49–59.

Mozharovskyi P (2015). *Contributions to Depth-based Classification and Computation of the Tukey Depth*. Ph.D. thesis.

Mozharovskyi P, Mosler K, Lange T (2015). "Classifying Real-world Data with the {DD}\ Alpha-procedure." *Advances in Data Analysis and Classification*, **9**(3), 287–314.

Paindaveine D, Van Bever G (2013). "From Depth to Local Depth: A Focus on Centrality." *Journal of the American Statistical Association*, **108**(503), 1105–1119.

Paindaveine D, Van Bever G (2015). "Nonparametrically Consistent Depth-based Classifiers." *Bernoulli*, **21**(1), 62–82.

Pokotylo O, Mosler K (2016). "Classification with the Pot-pot Plot." *arXiv preprint arXiv:1608.02861*.

Pokotylo O, Mozharovskyi P, Dyckerhoff R (2016). "Depth and Depth-based Classification with R-package ddalpha." *arXiv preprint arXiv:1608.04109*.

Vencalek O (2011). *Weighted Data Depth and Depth Based Discrimination*. Ph.D. thesis, Doctoral Thesis. Charles University. Prague. URL http://artax.karlin.mff.cuni.cz/~venco2am/DataDepth.html.

Vencalek O (2013). "New Depth-based Modification of the k-nearest Neighbour Method." *SOP Transactions on Statistics and Analysis*, **1**(2), 131–138.

Affiliation:

Ondřej Vencálek
Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University in Olomouc
17. listopadu 12, 771 46 Olomouc, Czech Republic
E-mail: ondrej.vencalek@upol.cz

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 46
April 2017

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2016-11-15
Accepted: 2017-02-02