Kuwait J.Sci., Vol.48 (2), April.2021, pp(1-11)

Depth classification based on affine-invariant, weighted and kernel-based spatial depth functions

Olusola S. Makinde

Dept. of Statistics, Federal University of Technology, P.M.B 704, Akure, Nigeria

Corresponding author: osmakinde@futa.edu.ng

Abstract

Several multivariate depth functions have been proposed in the literature, of which some satisfy all the conditions for statistical depth functions while some do not. Spatial depth is known to be invariant to spherical and shift transformations. In this paper, the possibility of using different versions of spatial depth in classification is considered. The covariance-adjusted, weighted, and kernel-based versions of spatial depth functions are presented to classify multivariate outcomes. We extend the maximal depth classification notions for the covariance-adjusted, weighted, and kernel-based spatial depth versions. The classifiers' performance is considered and compared with some existing classification methods using simulated and real datasets.

Keywords: Classifiers; data depth; kernel; spatial depth; weight.

1. Introduction

In classifying objects in $\mathbb{R}^d$, several classification methods in the literature include discriminant analysis, support vector machine (SVM) (Vapnik, 1998), among others. Some of these methods have intuitive features like robustness against outliers, optimality under special conditions for some distributions. However, these methods are either based on some distribution assumptions or assume some parametric surfaces (Makinde & Chakraborty, 2015). Some involve estimating location and scale parameters whose model estimates are affected by outlying observations if present in the data. However, for discriminant analyses, Hubert & Van Driessen, 2004, presented robust versions of the classification rules. Also, high dimension poses a significant challenge on many parametric classification methods.

Nonparametric analysis of multivariate outcomes based on data depth has received much attention over two decades. Liu *et al*., 1999, proposed various ideas on analysing multivariate outcomes based on data depth notions. Data depth measures how central a multivariate object is concerning an underlying multivariate distribution or data cloud. Using this notion, Tukey, 1975; Liu, 1990; Donoho & Gasko, 1992, defined several depth functions for multivariate outcomes. Jornsten, 2004, applied the notions of data depth to classification. Also, Ghosh & Chaudhuri, 2005, proposed some versions of the maximum depth classifier. The maximum depth classifier assigns an observation to the population or sample for which the classifier attains its highest depth value. Li *et al*., 2012;

Lange *et al*., 2014, proposed graphical approaches to classification based on depth-depth plot, a two-dimensional representation of multivariate objects by their data depths concerning known classes. Makinde & Chakraborty, 2018, proposed classifiers based on multivariate rank and rank regions. Makinde & Chakraborty, 2015; Makinde, 2020, proposed classifiers based on the distribution of versions of multivariate rank functions. In another development, Makinde, 2020, presented some distance-based classification methods for gene expression data.

Spatial depth function has been applied in classification and performs competitively; see Ghosh & Chaudhuri, 2005, for example. However, the maximum depth classification method based on spatial depth suffers from a lack of robustness of spatial depth function against affine invariance. Also, each observation contributes equal weight to the spatial depth value. In real applications, some data points may be more important than others. Instead of each observation contributing equally to the spatial depth, some observations contribute more weight than others. Based on these, covariance-adjusted, weighted, and kernel-based spatial depth functions are considered in this paper. Classification rules based on these versions of spatial depth functions are proposed. We compare the proposed classification rules' performance with some existing methods such as linear discriminant analysis, support vector machine, and maximum depth classification method based on projection depth.

2. Classification rule based on spatial depth

2.1 Spatial depth classifier

Suppose $\boldsymbol{Y}$ is a d-dimensional random vector having a distribution F. The spatial depth function of any point $\boldsymbol{x} \in \mathbb{R}^d$ concerning F is defined as

$$SD(\boldsymbol{x}, F) = 1 - \left\| E\left[\frac{\boldsymbol{x}-\boldsymbol{Y}}{||\boldsymbol{x}-\boldsymbol{Y}||}\right]\right\| \qquad (1)$$

where $||.||$ is the usual Euclidean norm. Spatial depth possesses some intuitive features. $SD(\boldsymbol{x}, F)$ characterizes the distribution F in the sense that a more considerable depth value indicates more central observation and a smaller depth value indicates extreme observation.

Let $\pi_1, \pi_2, \ldots, \pi_J$ be $J(\geq 2)$ populations having d-dimensional distribution functions $F_1, F_2, \ldots, F_J$. The classification rule for any $\boldsymbol{x} \in \mathbb{R}^d$ is defined as

$$\text{assign } \boldsymbol{x} \text{ to } \pi_k \text{ if } SD(\boldsymbol{x}, F_k) = \max_{1\leq j\leq J} SD(\boldsymbol{x}, F_j) \qquad (2)$$

In practice, $SD(\boldsymbol{x}, F_j)$ will hardly be known, and we need to estimate it from the training sample. Let $\boldsymbol{X}_{j1}, \boldsymbol{X}_{j2}, \ldots, \boldsymbol{X}_{jn_j} \in \mathbb{R}^d$ be a random sample from the population π_j having distribution F_j. We define the empirical depth functions as

$$SD(\boldsymbol{x}, \hat{F}_j) = 1 - \left\| \frac{1}{n_j}\sum_{i=1}^{n_j}\left[\frac{\boldsymbol{x}-\boldsymbol{X}_{ji}}{||\boldsymbol{x}-\boldsymbol{X}_{ji}||}\right]\right\|$$

for any $\boldsymbol{x} \in \mathbb{R}^d$, where $\hat{F}_j$ is the empirical distribution of the sample from F_j. Then the empirical classification rule for any $\boldsymbol{x} \in \mathbb{R}^d$ can be defined as

$$\text{assign } \boldsymbol{x} \text{ to } \pi_k \text{ if } SD(\boldsymbol{x}, \hat{F}_k) = \max_{1\leq j\leq J} SD(\boldsymbol{x}, \hat{F}_j).$$

A related depth function to $SD(\boldsymbol{x}, F_j)$ was proposed in Gao, 2003. Denote the depth function in Gao, 2003, by $SD_{Gao}(\boldsymbol{x}, F_j)$ for j-th distribution, then

$$SD_{Gao}(\boldsymbol{x}, F_j) = 1 - O(\boldsymbol{x}, F_j)$$

Where,

$$O(\boldsymbol{x}, F_j) = \left\| E\left[\frac{\boldsymbol{x}-\boldsymbol{X}_j}{||\boldsymbol{x}-\boldsymbol{X}_j||}\right]\right\|^2.$$

The relationship between the depth functions $SD(\boldsymbol{x}, F_j)$ and $SD_{Gao}(\boldsymbol{x}, F_j)$ can be expressed as

$$SD_{Gao}(\boldsymbol{x}, F_j) = SD(\boldsymbol{x}, F_j)\left(2 - SD(\boldsymbol{x}, F_j)\right).$$

It follows that $SD(\boldsymbol{x}, F_j)$ and $SD_{Gao}(\boldsymbol{x}, F_j)$ perform equivalently in maximal depth classification.

Zuo & Serfling, 2000, set out four conditions for statistical depth functions. Both $SD(\boldsymbol{x}, F_j)$ and $SD_{Gao}(\boldsymbol{x}, F_j)$ satisfy all the conditions set out for statistical depth function of spherical symmetric population. However, neither $SD_{Gao}(\boldsymbol{x}, F_j)$ nor $SD(\boldsymbol{x}, F_j)$ is invariant under general affine transformation, and hence classification rule based on any of them is non-invariant under the general affine transformation of the data.

Any invariant depth to shifts and orthogonal transformations can be affine invariant by sphering the data (transforming the depth by a proper matrix). There are many ways to do this. An example is to consider the root of the covariance matrix. Let $SD_a(\boldsymbol{x}, F_j)$ denote the spatial depth function based on F_j and $\boldsymbol{\Sigma}_j$, where, the d-dimensional random vectors $\boldsymbol{X}_j$ have distributions F_j for $j = 1,2,\dots,J$ and $\boldsymbol{\Sigma}_j$ is the covariance matrix of F_j,

$$SD_a(\boldsymbol{x}, F_j) = 1 - \left\| E\left[\frac{\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{X}_j)}{||\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{X}_j)||}\right]\right\|$$

where $\boldsymbol{A}$ is a $d \times d$ matrix such that $\boldsymbol{A}\boldsymbol{A}' = c\boldsymbol{\Sigma}$ for some constant c. If the covariance of the distribution F_j exists, we can take $\boldsymbol{A}$ to be the Cholesky decomposition of the covariance matrix.

The classification rule based on $SD_a(\boldsymbol{x}, F_j)$ is

$$\text{assign } \boldsymbol{x} \text{ to } \pi_k \text{ if } SD_a(\boldsymbol{x}, F_k) = \max_{1\le j\le J} SD_a(\boldsymbol{x}, F_j). \quad (3)$$

Where, there is no confusion, we denote the classification rule based on $SD(\boldsymbol{x}, F_j)$ in (2) and $SD_a(\boldsymbol{x}, F_j)$ in (3) by SD and SD_a respectively in our numerical examples.

2.2 Weighted spatial depth classifier

The weighted spatial depth function, denoted by $SD_w(\boldsymbol{x}, F_j)$ for population π_j with distribution F_j, can be defined as

$$SD_w(\boldsymbol{x}, F_j) = 1 - \left\| E[\omega_j S(\boldsymbol{X}_j, \boldsymbol{x})] \right\|$$

where

$$S(\boldsymbol{X}_j, \boldsymbol{x}) = \frac{\boldsymbol{x}-\boldsymbol{X}_j}{||\boldsymbol{x}-\boldsymbol{X}_j||}$$

and ω_j is the random weight for j population. Using $SD_w(\boldsymbol{x}, F_j)$ in place of $SD(\boldsymbol{x}, F_j)$ in (2), the resulting classification rule, denoted by S.D., is

$$\text{assign } \boldsymbol{x} \text{ to } \pi_k \text{ if } SD_w(\boldsymbol{x}, F_k) = \max_{1\le j\le J} SD_w(\boldsymbol{x}, F_j) \quad (4)$$

The sample version of $SD_w(\boldsymbol{x}, F_j)$ is defined as

$$SD_w(\boldsymbol{x}, \hat{F}_j) = 1 - \left\| \frac{1}{n_j}\sum_{i=1}^{n_j}\left[\omega_{ji}\frac{\boldsymbol{x}-\boldsymbol{X}_{ji}}{||\boldsymbol{x}-\boldsymbol{X}_{ji}||}\right]\right\|,$$

Where, weights $\omega_{ji}, \omega_{j2}, \dots, \omega_{jn_j}$ are weights of sample points $\boldsymbol{X}_{j1}, \boldsymbol{X}_{j2}, \dots, \boldsymbol{X}_{jn_j} \in \mathbb{R}^d$ from the population π_j and are non-negative for $j = 1,2,\dots,J$.

To implement the classification rule based on weighted spatial depth, the Gaussian weight function is used. The Gaussian weight employed for ith observation from jth class of size n_j is defined as

$$\omega_{ji} = r_{ji} / \sum_{i=1}^{n_j} r_{ji}$$

Where,

$r_{ji} = K(\boldsymbol{X}_i, \boldsymbol{Y}_j) = exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2\right)$, σ^2 is the turning parameter. The value of σ^2 is taken to be 9 for weighted spatial depth classifier in the numerical examples in the next section following Chen *et al*., 2009, argued that $K(\boldsymbol{X}_i, \boldsymbol{Y}_j)$ captures the shape of some multivariate data at $\sigma^2 = 9$.

2.3 Kernelized spatial depth classifier

Chen *et al*., 2009, defined a kernel-based spatial depth function and proposed an outlier detection methodology based on the depth function. The idea is to evaluate the spatial depth in a feature space induced by a definite positive kernel.

The kernelized spatial depth (Chen *et al*., 2009) is defined as

$$SD_k(\boldsymbol{x}, F_j) = 1 - O_k(\boldsymbol{x}, F_j)$$

where

$$O_k(\boldsymbol{x}, F_j) = \frac{1}{n}\left(\sum_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d} \frac{K(\boldsymbol{x},\boldsymbol{x}) + K(\boldsymbol{y},\boldsymbol{z}) - K(\boldsymbol{x},\boldsymbol{y}) + K(\boldsymbol{x},\boldsymbol{z})}{\zeta(\boldsymbol{x},\boldsymbol{y})\zeta(\boldsymbol{x},\boldsymbol{z})}\right)^{1/2},$$

$$\zeta(\boldsymbol{x},\boldsymbol{y}) = \sqrt{K(\boldsymbol{x},\boldsymbol{x}) + K(\boldsymbol{y},\boldsymbol{y}) - 2K(\boldsymbol{x},\boldsymbol{y})},$$

$$\zeta(\boldsymbol{x},\boldsymbol{z}) = \sqrt{K(\boldsymbol{x},\boldsymbol{x}) + K(\boldsymbol{z},\boldsymbol{z}) - 2K(\boldsymbol{x},\boldsymbol{z})},$$

And K is a kernel function based on semi-metric between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. The classification rule based on $SD_k(\boldsymbol{x}, F_j)$ is defined as

$$\text{assign } \boldsymbol{x} \text{ to } \pi_k \text{ if } SD_k(\boldsymbol{x}, F_k) = \max_{1\le j\le J} SD_k(\boldsymbol{x}, F_j). \quad (5)$$

The kernel functions in the literature include Gaussian kernel, polynomial kernel, Laplacian kernel, exponential kernel, among others. Gaussian Kernel, denoted by $K(\boldsymbol{x}, \boldsymbol{y})$, is defined as $K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{1}{h}\|\boldsymbol{x} - \boldsymbol{y}\|^2\right)$, where h is the tuning parameter. The rational quadratic kernel is computationally simple compared to the Gaussian kernel.
It is defined as

$$K(\boldsymbol{x}, \boldsymbol{y}) = 1 - \left(\frac{\|\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2}{c + \|\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2}\right),$$

where c is a constant term.

3. Simulation study

Simulation 1: Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_m$ are random training samples from trivariate normal populations π_1 and π_2 with center of symmetries $\boldsymbol{\mu}_1 = (1 \quad 0 \quad 0)'$ and $\boldsymbol{\mu}_2 = (1 \quad 1 \quad 1)'$, respectively and covariance matrix $\boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix.

Simulation 2: Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_m$ are normally distributed with means $\boldsymbol{\mu}_1 = (1 \quad 1 \quad 1)'$ and $\boldsymbol{\mu}_2 = (1 \quad 0 \quad 0)'$ respectively and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.3 & 0.5 \\ 0.3 & 1 & 0.4 \\ 0.5 & 0.4 & 1 \end{pmatrix}$.

Simulation 3: Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ is a random training sample from a standard mixture distribution defined as

$$F = \begin{cases} N(\boldsymbol{\mu}_X^1, \boldsymbol{I}) & \text{with probability } p \\ N(\boldsymbol{\mu}_X^2, \boldsymbol{I}) & \text{with probability } 1-p \end{cases}$$

where $p \in (0,1)$ is the mixing proportion, $\boldsymbol{\mu}_X^1 = (0 \quad 0)'$ and $\boldsymbol{\mu}_X^2 = (1 \quad 0)'$. Suppose $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_m$ is a random training sample from a bivariate normal distribution with mean vector $(2 \quad 0)'$ and covariance matrix $\boldsymbol{I}$.

Simulation 4: Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_{0.9n}$ and $\boldsymbol{X}_{0.9n+1}, \boldsymbol{X}_{0.9n+2}, \dots, \boldsymbol{X}_n$ are random training samples from trivariate independent exponential distributions $F = (exp(1) + 1, exp(1), exp(1))$ and $G = (exp(1) + 1, exp(1) + 1, exp(1) + 2)$, respectively. Suppose $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_m$ is a random training sample from trivariate independent exponential distribution $G = (exp(1) + 1, exp(1) + 1, exp(1) + 2)$.

The sample sizes for $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$ are taken to be $n = m = 20, 50, 100$. We simulate random from the distribution of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$ for each of the Simulations 1 - 4, then assign observations in the test samples to one of F and G. Setting $s = n = m$, the probability of misclassification is estimated by the mean of misclassification error rates in $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{2s}$. The simulation procedure is repeated 100 times.

Table 1. Effect of bandwidth on Gaussian kernel-based spatial depth classifiers (SDk) based on misclassification rates.

	values of bandwidth (h)			
Simulation	1	5	9	25
1	0.2357	0.2028	0.2094	0.2111
2	0.3018	0.2670	0.2639	0.2767
3	0.2487	0.2445	0.2318	0.2390
4	0.2360	0.2599	0.2674	0.2838

The kernelized spatial depth classifier using Gaussian kernel is denoted by SDk and kernelized spatial depth classifier using rational quadratic kernel by SDrq where there is no confusion. For SVM, C-support vector classification with Gaussian RBF-kernel was used. The default choice of parameters in the R package *kernlab* was used with Gaussian kernel, and 5-fold cross-validation was employed.

Simulation 1 consists of two spherically symmetric normal distributions with different location vectors. Simulation 2 consists of two elliptically symmetric normal distributions with different location vectors. Simulation 3 consists of spherically symmetric normal distribution and mixture normal distribution. Simulation 4 comprises two trivariate independent exponential samples, containing 10% contamination from another independent exponential distribution. test samples $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_s$ from the distribution of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Z}_{s+1}, \boldsymbol{Z}_{s+2}, \ldots, \boldsymbol{Z}_{2s}$

Table 3 presents averages of misclassification error rates of competing classifiers for spherically symmetric normal distribution (Simulation 1), elliptically symmetric normal distribution (Simulation 2), mixture normal distribution (Simulation 3), and independent exponential distribution with 10% contamination (Simulation 4) with different sample sizes.

The choice of bandwidth for SDk and SDrq were investigated in Simulations 1 – 4 for $m = n = 50$. Table 1 presents the performance of SDk for some values of bandwidth h. In Simulation 1, the average misclassification rate is least at $h = 5$. The average misclassification rates are least at the value of $h = 9$ in Simulations 2 and 3, where the competing distributions are elliptical and mixture distributions, respectively. This is in agreement with the claim of Chen *et al*., 2009. This may be attributed to the fact that SDk captures the shape of some multivariate data at $h = 9$. However, the average misclassification rate in Simulation 4 is minimized as lower value bandwidth($h = 1$).

Also, Table 2 presents the performance of SDrq for some values of bandwidth c. In Simulations 1, 3, and 4, the average misclassification rates are minimized at $c = 1$. The average misclassification rate in Simulation 2 is minimized at $c = 1.5$. For the implementation of SDk and SDrq, the choice of value for each of h and c were taken to be the optimal value obtained in Tables 1 - 2 in each simulation experiment.

In Table 3, a comparison of spatial depth classifiers (SD, SDw, SDk, and SDrq) was presented with some existing classification methods such as Fisher's linear discriminant

Table 2. Effect of bandwidth on rational quadratic kernel-based spatial depth classifiers (SDrq) based on misclassification rates.

Simulation	various values of c				
	0.1	0.5	1	1.5	2
1	0.2240	0.2225	0.2058	0.2158	0.2035
2	0.2823	0.2810	0.2750	0.2644	0.2657
3	0.2610	0.2491	0.2398	0.2473	0.2441
4	0.1445	0.1311	0.1250	0.1252	0.1285

analysis (LDA), maximal depth classifier based on projection depth (Donoho & Gasko, 1992) (P.D.) and SVM based on misclassification rates. For standard samples (Simulations 1 and 2), LDA performs best. It is observed that LDA, SD, and SDw perform equivalently for standard samples for different training sample sizes. The performance of SDa is better than that of S.D. in Simulation 2 because of the ability of SDa to take care of correlation among variables but for large sample sizes. In Simulation 3, all the competing classification methods perform almost equivalently for a large sample size. It is expected that LDA will outperform other methods in Simulation 1-3. This is because LDA is Bayes rule when data is normally distributed (either spherically or elliptically symmetric) (Anderson, 1984). It is noted that if the covariance matrix in Simulation 2 has more extreme entries, for example, correlation values among some variables increase, the average misclassification error rate of LDA reduces as some correlation values among variables increase (Fan *et al*., 2012). Similarly, the average misclassification error rates of SDa reduce as some correlation values among variables increase. For instance, suppose $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.85 \\ 0.5 & 0.85 & 1 \end{pmatrix}$, the average misclassification error rates of LDA, SDa, PD, and SVM reduce compared to the results in Simulation 2 for training sample sizes n=20, 50, 100. However, the average misclassification error rates of SD, SDw, SDk, and SDrq are equivalent to Simulation 2. This can be attributed to the lack of affine invariance of classifiers SD, SDw, SDk, and SDrq.

For exponential samples in Simulation 4, SVM performs best while SDrq and SDk compete well. SVM achieves the least average misclassification rate. This can be attributed to its optimal behaviour when competing populations are exponentially distributed. It is observed that SDk and SDrq perform competitively with LDA and P.D. for exponential samples. Figure 1 presents the boxplots of misclassification error rates of competing classifiers when $n = m = 100$ for the four simulation examples.

It can be inferred from Table 3 that S.D. and SDw can be used as alternative classification methods to LDA for spherically symmetric normally distributed data. These classifiers are robust against outliers if present in the data while LDA does not. The classifier SDa can be employed as an alternative LDA for elliptically symmetric normally distributed data. The classifiers SD, SDw, and P.D., can be used instead of LDA for normal mixture distribution. For exponentially distributed data, SVM is optimal. The spatial depth classifiers SDk and SDrq are preferred over LDA and P.D. The R codes for the spatial classification methods are freely available at https://github.com/osMakinde/osMakinde_spatial_depth_classifiers.

Table 3. Average misclassification error rates of competing classifiers for trivariate normal and exponential samples with varying sample sizes.

Simulation	Sample size	LDA	PD	SVM	SD	SDa	SDw	SDk	SDrq
1	20	0.206	0.269	0.301	0.222	0.238	0.209	0.220	0.243
	50	0.198	0.223	0.284	0.200	0.211	0.203	0.200	0.212
	100	0.196	0.210	0.275	0.196	0.201	0.196	0.201	0.207
2	20	0.269	0.325	0.308	0.284	0.306	0.289	0.304	0.305
	50	0.250	0.281	0.273	0.271	0.265	0.278	0.271	0.264
	100	0.252	0.269	0.257	0.265	0.250	0.271	0.265	0.270
3	20	0.235	0.270	0.266	0.252	0.257	0.245	0.249	0.269
	50	0.236	0.244	0.247	0.234	0.230	0.231	0.233	0.240
	100	0.225	0.228	0.232	0.229	0.236	0.228	0.239	0.233
4	20	0.301	0.325	0.235	0.321	0.334	0.304	0.267	0.257
	50	0.295	0.301	0.194	0.303	0.314	0.315	0.234	0.231
	100	0.291	0.292	0.171	0.294	0.300	0.293	0.212	0.210

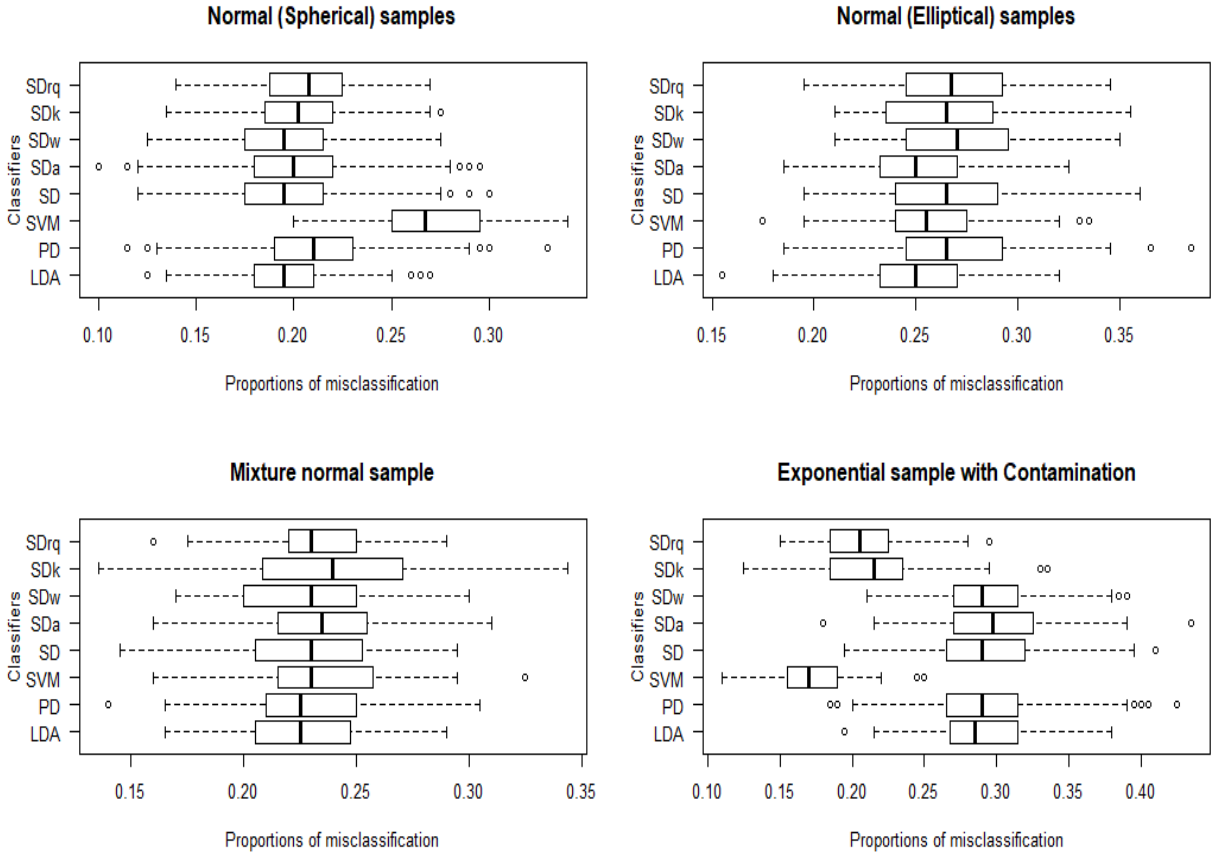

Fig. 1. Boxplots of misclassification error rates of competing classifiers

4. Real data

Four benchmark data sets are analysed to illustrate the performances of the proposed spatial depth classification methods. These datasets are South African heart disease data, *E coli* data, Iris data, and glass data. South African heart disease data, denoted by SAHD, consists of two classes (response and coronary heart disease). A random training and test sample of size 100 and 40 respectively are chosen from each of the two classes with eight input features. SAHD data is available at http://www.stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data. *E. coli* data consists of two classes with 5 features. Ecoli data and its description are available at https://archive.ics.uci.edu/ml/machine-learning-databases/ecoli/ecoli.data.
A random training sample of size 50 and a test sample of size 25 is chosen from each of the classes. Iris data consists of three classes (Setosa, Versicolor, and Virginica). A random training and test sample of size 25 is chosen from each of the three classes. Glass data consists of seven classes and 9 input features. A random training sample of size 50 is chosen from each of the first two classes, while the test samples consist of a complement of the training samples. We take the bandwidth h for SDk to be 5.0 for *E. coli* and iris data while $h = 1$ for glass data and $h = 0.25$ for South African heart disease data. Also, the bandwidth c for SDrq is chosen to be 1.0 for the four datasets.

Table 4 presents the average misclassification error rates of SD, SDa, SDw, K.S.D., and SDrq compared to LDA, P.D., and SVM. For SAHD, SDk achieves perfect classification. Other competing classification methods perform equivalently. For *E. coli* data, LDA and SVM have the least average error rate while SDa performs competitively. For iris data, SDa performs best while P.D., SVM, SD, SDw, and SDk compete well. LDA has the highest average error rates. For glass data, SDrq, SVM, and SDk have the least average misclassification error rates while SDa competes well with LDA. The classifiers, S.D. and SDw, have the highest average misclassification error rates. Figure 2 presents boxplots of classifiers' performance based on average misclassification error rates for real data sets.

It has been discussed in the literature that iris data is normally distributed for each species. Makinde & Chakraborty, 2018, have shown that classification methods based on the function of spatial rank outlyingness are a Bayes rule for elliptically symmetric normally distributed multivariate. Under this condition, SDa is a Bayes rule. Comparing spatial depth classifiers, it is observed that SDw, SDk, and SDrq perform equivalently for *E. coli* and iris data. For all the datasets, spatial depth classifiers S.D. and SDw perform equivalently.

Table 4. Average error rates of classifiers for some real datasets.

Dataset	LDA	PD	SVM	SD	SDa	SDw	SDk	SDrq
SAHD	0.313	0.341	0.501	0.330	0.348	0.332	0.000	0.343
Ecoli	0.026	0.047	0.026	0.058	0.039	0.058	0.057	0.058
Iris	0.164	0.042	0.052	0.065	0.029	0.065	0.059	0.064
Glass	0.294	0.326	0.214	0.396	0.290	0.399	0.224	0.207

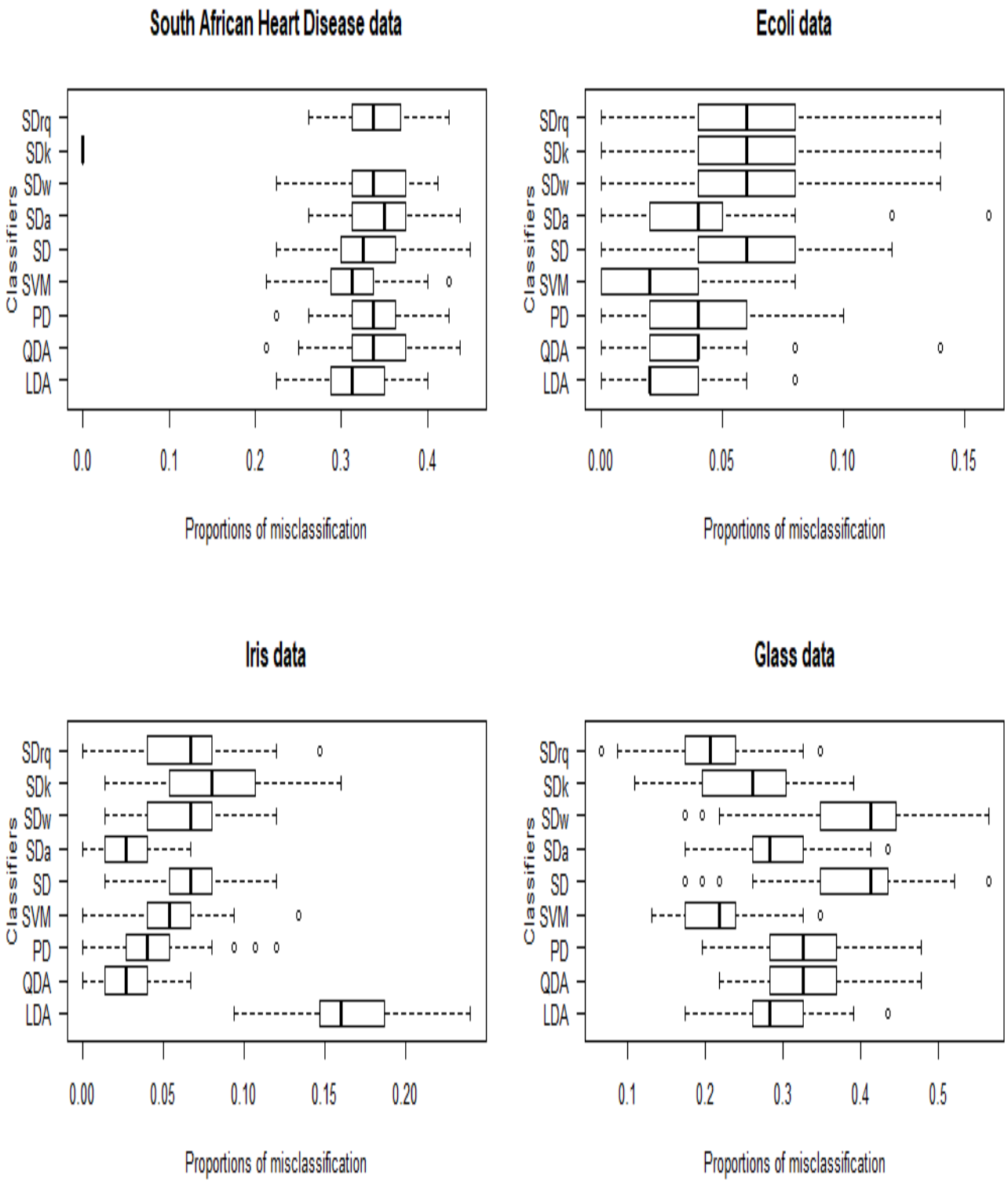

Fig. 2. Boxplots of misclassification error rates of competing classifiers for real data.

5. Summary

Some forms of spatial depth function are defined in this study. A big advantage of the spatial depth over other notions of depth (such as location or projection depth) is its computational simplicity, which is linear both in dimension and in sample length. These depth functions are considered to classify observations in test set to one of the known classes correctly using maximum depth classification rules of Ghosh & Chaudhuri, 2005. The depth functions are spatial depth, a modified version based on the cholesky decomposition of the data cloud's covariance matrix, weighted spatial depth, and some kernelized spatial depths. The performance of each associated classifier is examined using simulation as well as real data. Based on results from simulation and real data examples, classifiers based on some forms of spatial depth function, for example, SD, SDa, and SDk, perform well and competitively with others. When there is evidence of correlation among the data features, the maximum depth classification rule's better performance is observed with a modified spatial depth version (SDa). When some data points are more important than the other, weighted and kernelized spatial depth functions are recommended to be employed for maximum depth classification. For kernelized spatial depth function, it is observed that the impact of an optimal choice of bandwidth is more than the impact of the kernel in the classification.

References

Anderson, T. W. (1984) An introduction to multivariate statistical analysis. NY: John Wiley & Sons, Inc, **207-250**.

Chen, Y., Dang, X., Peng, H. & Bart, H.L. (2009) Outlier detection with the kernelized spatial depth function. IEEE Transactions on Pattern Analysis and Machine Intelligence. **31:288-305**.

Donoho, D.L. & Gasko, M. (1992) Breakdown properties of multivariate location parameters and dispersion matrices, Annals of Statistics, **20:1803-1827.**

Fan, J., Feng, Y. & Tong, X. (2012) A road to classification in high dimensional space: the regularized optimal affine discriminant. Journal of the Royal Statistical Society: Series B, **74(4):745-771.**

Gao, Y. (2003) Data depth based on spatial rank. Statistics & Probability Letters, **65:217-225**.

Ghosh, A.K. & Chaudhuri, P. (2005) On maximum depth and related classifiers. Scandinavian Journal of Statistics, **32:327-350.**

Hubert, M. & Van-Driessen, K. (2004) Fast and robust discriminant analysis. Computational Statistics and Data Analysis, **45(2):301-320.**

Jornsten, R. (2004) Clustering and classification based on the L_1 data depth. Journal of Multivariate Analysis, **90:67-89.**

Li, J., Cuesta-Albertos, J.A. & Liu, R.Y. (2012) DD-classifier: nonparametric classification procedure based on DD-plot. Journal of American Statistical Association, **107:737-753**.

Liu, R.Y. (1990) On a notion of data depth based on random simplices. Annals of Statistics, **18:405-414.**

Liu, R.Y., Parelius, J.M. & Singh, K. (1999) Multivariate analysis by data depth:

Descriptive statistics, graphics, and inference. Annals of Statistics, **27:783-858**.

Makinde, O.S. & Chakraborty, B. (2015) On some nonparametric classifiers based on distribution functions of multivariate ranks. In Nordhausen, K and Taskinen, S.(eds): Modern Nonparametric, Robust and Multivariate Methods, Festschrift in Honour of Hannu Oja. Springer, Switzerland, **249-264**.

Makinde, O.S. & Chakraborty, B. (2018) On some classifiers based on multivariate ranks. Communication in Statistics - Theory and Methods, **47(16):3955-3969. DOI:10.1080/03610926.2017.1366520**

Makinde, O.S. (2019) Classification of gene expression data: Distance-based method. Kuwait Journal of Science, **46(3):31-39**.

Makinde, O.S. (2020) On rank distribution classifiers for high dimensional data. Journal of Applied Statistics, **47:2895--2911.**

Tukey, J. (1975) Mathematics and picturing data. Proceedings of the 1975 International Congress of Mathematics, **2:523-531**.

Vapnik, V.N. (1998) Statistical Learning Theory. John Wiley and Sons, New York, **375-570**.

Zuo, Y. & Serfling, R. (2000) General notions of statistical depth function. Annals of Statistics, **28(2):461-482**.

Submitted **:** 26/11/2019

Revised **:** 25/05/2020

Accepted **:** 11/06/2020

DOI **:** 10.48129/kjs.v48i2.8693