Classification of High Dimensional Data: A Distance Based Approach

Olusola Samuel Makinde[a,b] and Biman Chakraborty[b]
[a]Department of Statistics, Federal University of Technology, Akure, Nigeria
[b]School of Mathematics, University of Birmingham, Birmingham B15 2TT, United Kingdom
[a]osmakinde@futa.edu.ng, [b]B.Chakraborty@bham.ac.uk

Abstract

In this paper, a classification method based on L_2 distance of unclassified observations to spatial median of the data cloud for high dimensional data is proposed. Possibility of using depth oriented medians in place of spatial medians is also considered. The performance of the proposed method is examined by using simulations and the results are compared with the results from some existing methods. Analysis of real data examples shows that median based methods yield a good performance among their competitors.

Keywords: spatial median; depth oriented median; high dimension; classifier.

1 Introduction

In classifying objects in $\mathbb{R}^d$, classification rules based on distance measures perform well under suitable conditions and can be viewed as alternatives to popular methods in literature such support vector machine (Vapnik, 1998), maximum depth classifiers (Ghosh and Chaudhuri, 2005) and logistic regression in low dimension setting. Distance based classification methods include discriminant analysis(Fisher, 1936), centroid based classifiers (Hastie et al., 2001), k-nearest neighbour rule, among others. Some of these methods have intuitive features like optimality under necessary conditions. However, these methods are either based on some distribution assumptions or assume some parametric surfaces. Some involve estimating location and scale parameter whose model estimates are affected with outlying observations if present in the data.

When dimension of data cloud is greater than the sample size, implementation of many of the classification methods becomes practically difficult especially for discriminant analysis and maximal depth classifiers(Ghosh and Chaudhuri, 2005). Hall et al. (2009) proposed a classification method based on minimising L_1 distance to component-wise median to solve classification problem in high dimension. This method performs well, especially when competing distributions are heavy-tailed. Makinde and Chakraborty (2015) proposed multivariate rank based classification methods.

Multivariate median is a nonparametric and robust estimate for the centre of multivariate distribution or data cloud. The multivariate medians include spatial median, componentwise median, depth oriented median, etc. In this paper, classification rule based on Euclidean distance of test observations to spatial median is proposed. Use of depth oriented medians, such as half-space median, random projection median, is raised. The performance of the proposed classifiers is compared with some existing classification methods using simulation and real data sets.

2 Classification rule

Suppose $\mathbf{X}$ is a d-dimensional random vector having a distribution F, which is assumed to be absolutely continuous with respect to the Lebesgue measure $\mathbb{R}^d$. The spatial median of $\mathbf{X} \in \mathbb{R}^d$ with respect to distribution F is defined as

$$\mathbf{m} = \arg\min_{\mathbf{y}} E[\|\mathbf{y} - \mathbf{X}\| - \|\mathbf{X}\|]$$

where $\|\cdot\|$ is the usual Euclidean norm. Alternatively, Makinde and Chakraborty (2015) defined a spatial median as a point in F whose spatial rank outlyingness is zero. Liu et al. (1999) defined a data depth as measure of how outlying or central an observation is with with respect to data cloud. It follows immediately that the spatial median of the distribution F is the point in $\mathbb{R}^d$ with highest spatial depth value.

Consider J populations $\pi_1, \pi_2, \ldots, \pi_J$, then assign $\mathbf{x}$ to population π_k with distribution F_k if

$$\mathcal{D}(\mathbf{x}, \mathbf{m}_k) = \min_{1 \le j \le J} \mathcal{D}(\mathbf{x}, \mathbf{m}_j), \qquad 1 \le j \le J$$

where $\mathcal{D}(\mathbf{x}, \mathbf{m}_j) = \|\mathbf{x} - \mathbf{m}_j\|$ and $\mathbf{m}_j$ is the spatial median of the distribution F_j.

In practice, $\mathbf{m}_j$ will hardly be known completely and we need to estimate them from the training samples. Let $\mathbf{X}_{j1}, \mathbf{X}_{j2}, \ldots, \mathbf{X}_{jn_j} \in \mathbb{R}^d$ be a random sample from the population π_j having distribution F_j. We define the empirical spatial median $\hat{\mathbf{m}}_j$ as

$$\hat{\mathbf{m}}_j = \arg\min_{\mathbf{y}} \sum_{i=1}^{n_j} [\|\mathbf{y} - \mathbf{X}_{ji}\| - \|\mathbf{X}_{ji}\|].$$

The empirical classification rule for any $\mathbf{y} \in \mathbb{R}^d$ can be defined as

$$\text{assign } \mathbf{y} \text{ to } \pi_k \text{ if } D(\mathbf{y}, \hat{\mathbf{m}}_k) = \min_{1 \le j \le J} D(\mathbf{y}, \hat{\mathbf{m}}_j). \tag{1}$$

We denote the classification rule in (1) by D-SM where there is no confusion.

It is observed that depth oriented median, observation with highest depth value in the data cloud, can be used in place of spatial median for the above classification rule depending on the notion of data depth considered. Depth oriented medians in literature include half-space median, simplicial median, projection median, among others. However most of the depth oriented medians are limited in application due to their computational difficulty especially for large dimension. Exact half-space median and simplicial median can only be computed when dimension of data cloud is 3 and 2 respectively. Projection median is the observation in the data cloud with highest depth value. It can be computed in high dimension. The classification rule based on projection median, denoted by D-RP, assigns an observation to the group for which it attains minimum L_2 distance to the group projection median.

3 Numerical Results

3.1 Simulation

Here, some simulation studies to investigate the performance of our proposed classification method in high dimension. Suppose there are 200 observations equally split between two competing groups G_1 and G_2. Each experiment consists of measurements on 1000 features with 50 observations belonging to each training set and 50 observations to each test set.

[Simulation 1] Suppose ith observation is in kth group, then $\mathbf{Y}_i \sim N(\boldsymbol{\mu}_k, \boldsymbol{I})$, where $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, \ldots, \mu_{k1000})^\top$ with $\mu_{1j} = 0$ for $1 \le j \le 1000$, $\mu_{2j} = 0.7$ if $1 \le j \le 500$ and $\mu_{2j} = 0$ otherwise, $\boldsymbol{I}$ is an identity matrix and $k = 1, 2$.

[Simulation 2] Suppose each experiment consists of measurements on independent features such that for $i \in G_1$, $Y_{ij} \sim exp(1)$ for $1 \le j \le 1000$ and for $i \in G_2$, $Y_{ij} \sim exp(1) + 1$ if $1 \le j \le 500$ and $Y_{ij} \sim exp(1)$ if $501 \le j \le 1000$.

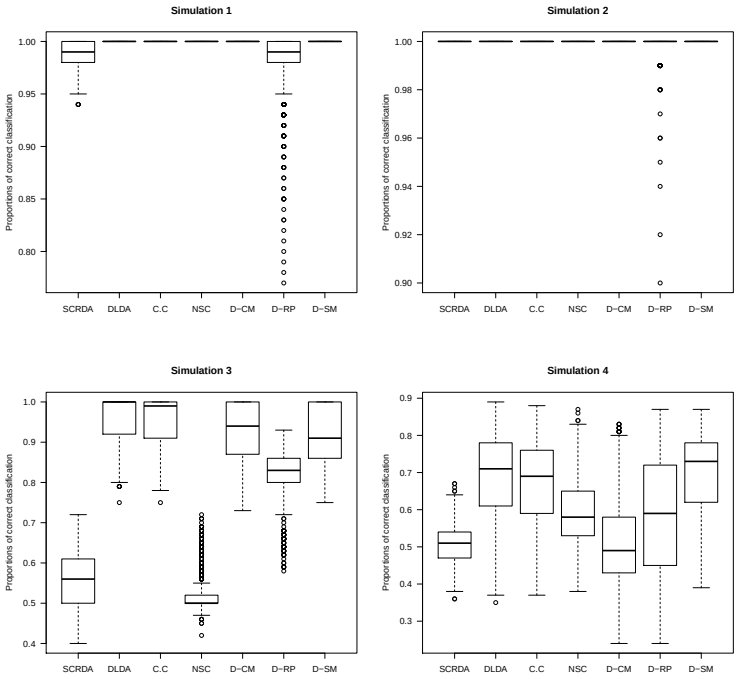

Figure 1: Box plot of proportions of correct classification for simulated data in high dimension.

[Simulation 3] Suppose the distribution F is normal mixture distributions, defined as

$$F = \begin{cases} N(\boldsymbol{\mu}_1^1, \boldsymbol{I}), \text{ with } p \\ N(\boldsymbol{\mu}_1^2, \boldsymbol{I}), \text{ with } 1-p \end{cases}$$

and the distribution G is multivariate normal distribution $N(\boldsymbol{\mu}_2, \boldsymbol{I})$, where $p \in (0,1)$ is the mixing proportion, $\mu_{1j}^1 = 0$ for $1 \leq j \leq 1000$, $\mu_{1j}^2 = 0.7$ if $1 \leq j \leq 500$ and $\mu_{1j}^2 = 0$ if $501 < j \leq 1000$ and $\boldsymbol{I}$ is an identity matrix .

[Simulation 4] Suppose G_1 consists of observations $\mathbf{Y}_i,\ i = 1, \ldots, 50$ such that

$$\mathbf{Y}_i = \begin{cases} \mathbf{Y}_i^0, \text{ with } p \\ \mathbf{Y}_i^1, \text{ with } 1-p \end{cases}$$

where $Y_{ij}^0 \sim exp(1)$ and $Y_{ij}^1 \sim exp(1) + 1$ for $1 \leq j \leq 1000$. Suppose G_2 consists of observations $\mathbf{X}_i = \{X_{i1}, X_{i2}, \cdots, X_{i1000}\}$, $i = 1, \ldots, 50$, where $X_{ij} \sim exp(1) + 0.5$ for $1 \leq j \leq 1000$.

The mixing proportion p is taken to be 0.7 and 0.6 for simulations 3 and 4 respectively. We compare the performance of the D-SM and D-RP with shrunken centroid regularized discriminant analysis (SCRDA)(Guo et al., 2007), diagonal linear discriminant analysis (DLDA), centroid classifier (C.C) (Hastie et al., 2001), nearest shrunken classifier (NSC) and componentwise median classifier (D-CM)(Hall et al., 2009). We have tuned the threshold parameter for NSC to be 1 and parameters α and δ of SCRDA to be 0.2 and 0.5 respectively.

Figure 1 presents the performance of the classifiers in terms of proportions of correct classification. All the competing classifiers achieve 100% proportions of correct classification in Simulation 1 except D-RP and SCRDA while all the classifiers perform well in Simulation 2. In Simulations 3 and 4, D-SM and D-CM perform competitively in terms of mean proportion of correct classification.

Small (1990) presented a survey of multidimensional medians. These include L_1 median, Oja simplex median, half-space median and simplicial median. These medians work well in low dimension setting. To

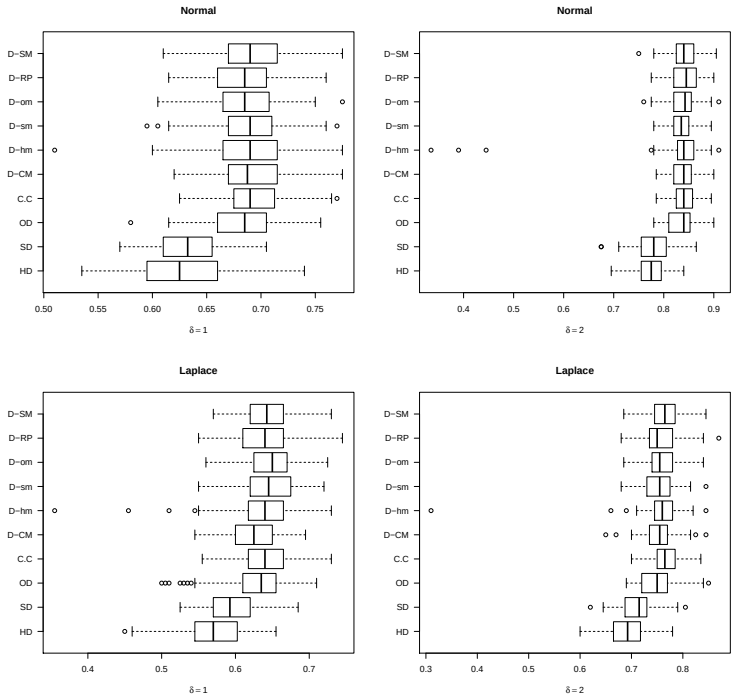

Figure 2: Box plots of proportions of correct classification for simulated data in low dimension.

illustrate the performance of these classifiers in low dimension, we present a simulation study. Consider F and G to be bivariate spherically symmetric distributions with centre of symmetries $\boldsymbol{\mu}_1 = (0,0)^\top$ and $\boldsymbol{\mu}_2 = (\delta,0)^\top$, respectively. Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ are random samples from F and G respectively, where n_1 and n_2 are taken to be 100. We simulate a new random sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$ from F and $\mathbf{Z}_{m+1}, \ldots, \mathbf{Z}_{2m}$ from G with $m = 100$ and compute the proportion of correct classification in $\mathbf{Z}_1, \ldots, \mathbf{Z}_{2m}$. The simulation size is 1000. For $\delta = 1, 2$, normal and Laplace samples are considered. These multivariate medians are considered in our simulated examples. D-SM is compared with L_2 distance to half-space median (D-hm), L_2 distance to simplicial median (D-sm), L_2 distance to Oja median (D-om), L_2 distance to projection median (D-RP), centroid classifier (Hastie et al., 2001) and maximal depth classifiers based on half-space depth(denoted by HD), simplicial depth (denoted by SD) and Oja depth (denoted by OD) using the simulation procedure described above.

Figure 2 presents the performance of competing classifiers in terms of proportions of correct classification. Bayes equivalence of maximal depth classifiers based on half-space depth(HD), simplicial depth(SD) and Oja depth (OD) was established in Ghosh and Chaudhuri (2005). Under independence of features and normality of competing classes, C.C is equivalent to Bayes rule. The fact that the mean proportion of correct classification of maximum depth classifiers and distance based classifiers are equivalent suggests that distance based classifiers are equivalent to Bayes rule. However, this claim needs theoretical verification and validity.

3.2 Real data

Two data sets are analysed to illustrate the performance of the proposed classification methods. The real datasets are lung cancer data and Leukaemia data, and are available in R package *rda*. Colon cancer data is a sparse data with two classes of sizes 22 and 40 with 2000 genes. Feature selection is performed on the colon cancer data to remove non-contributing genes using SCRDA with parameters $\alpha = 0.2$ and $\delta = 0.4$ as discussed in Guo et al. (2007). We choose a random training sample of size 15 and 30 while random test samples are taken to be the complementary of the training data. Leukaemia data consists of two groups

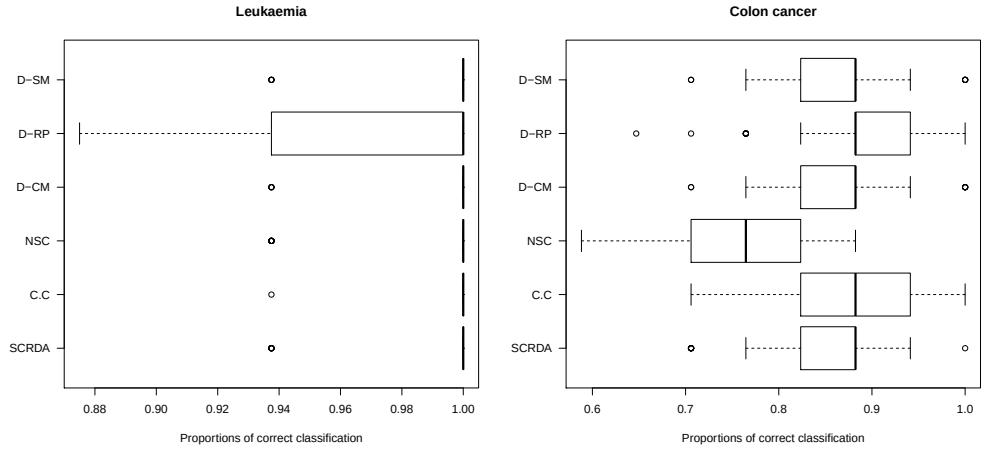

Figure 3: Box plot of proportions of correct classification for colon cancer and leukaemia data.

of sizes 27 and 11 with 3051 features. Feature selection is performed on the leukemia dataset to remove non-contributing genes using SCRDA with parameters $\alpha = 0.1$ and $\delta = 0.9$. Random training samples of sizes 15 and 7 are chosen and while random test samples are taken to be the complementary of the training data. The performances of D-SM and D-RP are compared with that of SCRDA, DLDA, NSC, D-CM and C.C. The choice of values of parameters α and δ for SCRDA classifier is as considered for feature selection.

Figure 3 present the proportions of correct classification of competing classifiers for colon cancer and leukaemia data using boxplots. All the classifiers perform well for leukaemia data. Averages of proportions of correct classification of SCRDA, C.C, NSC, D-CM, D-SM and D-RP(distance to projection median) are 0.9956, 0.9994, 0.994, 0.9975, 0.9975 and 0.9706 respectively. For colon cancer data, D-sm and D-RP compete well with their competitors, however D-RP performs best. Averages of proportions of correct classification of SCRDA, C.C, NSC, D-RP, D-CM and D-SM are 0.8518, 0.8682, 0.7506, 0.9047, 0.8641 and 0.8641 respectively.

4 Conclusion

Classification rule based on spatial median and generally, depth oriented median perform well when the dimension is less than sample size depending on the notion of data depth. However in high dimension, use of depth oriented median is limited due to computational difficulty. Classification method based on L_2 distance of test observation to spatial median and projection median can be computed for any dimension.

References

Fisher, R. A.(1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179–188.

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

Guo, Y., Hastie, T. and Tibshirani, R.(2007) Regularized linear discriminant analysis and its application in microarrays, Biostatistics, 8(1), 86–100

Hall P, Titterington DM, Xue J.(2009) Median based classifiers for high dimensional data, Journal of the American Statistical Association, 104, 1597–1608

Hastie, T., Tibshirani, R. and Friedman, J. H.(2001) The elements of statistical learning: data mining, inference and prediction. Springer-Verlag, New York.

Liu, R. Y., Parelius, J. M. and Singh, K.(1999) Multivariate analysis by data depth: Descriptive statistics, graphics and inference. The Annals of Statistics, **27**, 783–858.

Makinde, O. S. and 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, 249–264

Small, C.G. (1990). A survey of multidimensional medians. International Statistical review, 58(3):263–277.

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