Kuwait J. Sci. **46 (3)** pp. 31-39, 2019

Gene expression data classification: some distance-based methods

Olusola Samuel Makinde
Dept. of Statistics, Federal University of Technology, Akure, Nigeria
osmakinde@futa.edu.ng

Abstract

Micro-array dataset is a classical example of high throughput data characterized with more features (genes) than sample points (gene expression levels). A number of classification techniques have been proposed in literature. Many of these methods are either computationally expensive or perform sub-optimally. In this paper, some distance functions are considered and classification rules based on the distance functions are formulated. The distance functions include average distance measure, distance to component-wise median, distance to mean. We also define a probabilistic approach to classification rules based on two of the distance measures. Gene selection technique based on shrunken centroids regularized discriminant analysis was employed on small round blue cell tissue, colon cancer, lymphoma, prostate cancer and leukaemia data before applying the classification rules. Three simulation studies were performed to mimic gene expression data. The performance of the classification methods mentioned above was compared with performance of some known classification methods in literature. The performance of the distance-based classification methods is competitive with some existing classification methods. Distance based methods implemented in this study are computationally simple and very cheap in terms of computational cost.

Keywords: Distance methods; gene classifier; gene expression; proportion of correct classification.

1. Introduction

The objectives of classification as a practical subject in statistics are to find characteristics that define each competing classes and build a function or rule that assigns observations from unknown classes to one of competing classes on the basis of these characteristics (Dabney, 2005). A typical example is to distinguish between tumour tissues and normal tissues in colon tissue dataset (Alon *et al*., 1999). In micro-array data analysis, observations are referred to as gene expression levels. Each gene expression level consists of a number of genes. Micro array data is characterised by large number of genes but very few gene expression levels. Many of these genes are either irrelevant or redundant for discrimination among groups of gene expression data (Klassen and Kim, 2009).

Support vector machines have been employed in supervised statistical learning of gene expression data; see for example, Furey *et al*. (2000), Wang *et al*. (2008), Colak *et al*. (2016). Popularity of support vector machines can be attributed to its successful performance in many applications. However, redundant genes and extremely outlying genes can have serious impacts on support vector machines (Li and Yu, 2008). Vanitha *et al*. (2015) suggested selecting the informative genes based on the value of mutual information between the genes and the known gene classes.

Statistical distances are viable tools in the analysis of high dimensional data. Many statistical methodologies are based on these distances. Examples include k-means algorithms in cluster analysis, Hotelling T^2 test in statistical inference for location parameters, k-nearest neighbour rule in classification, Hotelling T^2 statistic in statistical quality control, among others. Hastie *et al*. (2001) proposed nearest centroid classifier for high dimensional data. Dabney (2005) applied the nearest centroid classifier to gene expression data with specific choice of active genes to participate in classification exercise. Tibshirani *et al*. (2002) proposed nearest shrunken centroid classifier (NSC), which is the modification of nearest centroid classifier. Klassen and Kim (2009) presented a combination of nearest shrunken centroid (NSC) as gene selection technique and random forest as a classifier for some gene expression data as well as its comparison with the performance of NSC as classifier.

Performances of classification methods, in general, can be evaluated based on their proportions of correct assignment or classification, interpretation of classification results and practical implementation of the classification methods in low and high dimensions. In this paper, we survey some existing distance functions and discuss some of their intuitive features such as computational simplicity. We define some classification rules based on these distance functions. The classification rules were employed on some gene expression data. For implementation of these classification approaches for gene expression data, gene selection technique based on shrunken centroid regularized discriminant analysis (Guo *et al.*, 2007) is employed to identify informative genes. We define distribution function of the distance measures discussed and formulate classification rules based on them. Also, we simulate data following some intuitive characteristics of gene expression data and employ the classification rules on the simulated data to evaluate their performance.

2. Methods

Suppose $\boldsymbol{X}_k \in \mathbb{R}^p$ is a random vector (or denotes random gene expression level in a micro-array experiment) having a distribution F_k, $k = 1, 2, \dots, K$. Define a distance measure of a vector $\boldsymbol{x} \in \mathbb{R}^p$ with respect to F_k as

$$D_1(\boldsymbol{x}, F_k) = E[||\boldsymbol{x} - \boldsymbol{X}_k||],$$

where$||.||$can be taken as the usual Euclidean norm and E denotes mathematical expectation. Suppose $F_1, F_2, \dots, F_K$ are distributions of competing classes $C_1, C_2, \dots, C_K$ respectively. We define the classification rule based on average distance as: assign a gene level z to class C if

$$D_1(\boldsymbol{x}, F_l) = \min_{1<k<K} D_1(\boldsymbol{x}, F_k). \qquad (1)$$

For latter reference, the classification rule in (1) is denoted by D_1 . The sample version of $D_1(\boldsymbol{x}, F_k)$ is defined thus; let $\boldsymbol{X}_{k1}, \boldsymbol{X}_{k2}, \dots, \boldsymbol{X}_{kn_k}$ be a random sample from F_k, one may define a distance function of a vector $\boldsymbol{x}$ based on $\boldsymbol{X}_{k1}, \boldsymbol{X}_{k2}, \dots, \boldsymbol{X}_{kn_k}$ as

$$D_1\left(\boldsymbol{x}, F_{n_k}\right) = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\sum_{j=1}^{p}\left(x_j - X_{kij}\right)^2\right)^{1/2}$$

$i = 1, 2, \dots, n_k$, $j = 1, 2, \dots, p$, $k = 1, 2, \dots, K$. This distance function $D_1(\boldsymbol{x}, F_k)$ is referred to as average distance function.

Another distance function in literature is L_2 distance to mean vector. This is defined as

$$D_2(\boldsymbol{x}, F_k) = ||\boldsymbol{\mu}_k - \boldsymbol{x}||,$$

where $\boldsymbol{\mu}_k$ denotes expected value of F_k. Hastie *et al.* (2001) presented this distance function in componentwise form and proposed a classification rule based on scaled version of $D_2(\boldsymbol{x}, F_k)$. It suffices to note that both $D_1(\boldsymbol{x}, F_k)$ and $D_2(\boldsymbol{x}, F_k)$ characterize F_k in the sense that high values of the distance metrics imply deviation from the centre of data cloud. It is worth mentioning that a generalized distance (also referred to as Mahalanobis distance) can be used instead of $D_2(\boldsymbol{x}, F_k)$. However, Mahalanobis distance is of limited use and particularly difficult to implement when number of features is greater than available sample size, for example in gene expression data. Suppose $F_1, F_2, \dots, F_K$ are distributions of completing classes $C_1, C_2, \dots, C_K$ respectively. We define the classification rule based on $D_2(\boldsymbol{x}, F_k)$ as: assign a gene level z to class C_l if

$$D_2(\boldsymbol{x}, F_l) = \min_{1<k<K} D_2(\boldsymbol{x}, F_k). \qquad (2)$$

For latter reference, the classification rule in (2) is denoted by D_2. The empirical version of $D_2(\boldsymbol{x}, F_k)$ is constructed by replacing population mean vector by sample mean vector. That is, the sample version of $D_2(\boldsymbol{x}, F_k)$ is defined as

$$D_2(\boldsymbol{x}, F_{n_k}) = \left[\sum_{j=1}^{p}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} X_{kij} - x_j\right)^2\right]^{1/2} = \left[\sum_{j=1}^{p}\left(\bar{X}_{kj} - x_j\right)^2\right]^{1/2}.$$

Convergence of $D_2\left(\boldsymbol{x}, F_{n_k}\right)$ to $D_2(\boldsymbol{x}, F_k)$ follows from the fact that $\bar{\boldsymbol{X}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{X}_{ki} \to \boldsymbol{\mu}$ for large sample sizes, $n_1, n_2, \dots., n_K$.

The classifier D_2 lacks robustness against outlying training sample points because a single outlying sample point can adversely affect sample mean vector, thereby affects the performance of the classifier D_2. In low dimensional setting, this problem can be easily overcome by using minimum covariance determinant (MCD) estimate of mean vector or trimmed mean vector. However, computation of minimum covariance determinant (MCD) estimate of mean vector is difficult in micro-array experiment because the ratio of number of genes to number of gene levels is far greater than 1. Depth oriented trimmed mean (Liu *et al.*, 1999) based on spatial depth can be employed.

When data cloud is heavy tailed, Hall *et al.* (2009) suggested use of L_1 distance to componentwise median.

The intuition behind it is that it is componentwise median that minimises L_1 distance among observations in the data cloud. L_1 distance to componentwise median is defined as

$$D_3(\boldsymbol{x}, F_k) = ||\boldsymbol{m}_k - \boldsymbol{x}||_1,$$

where $||\cdot||_1$ denotes L_1 norm and $\boldsymbol{m}_k$ is componentwise median of F_k. We define the classification rule based on $D_3(\boldsymbol{x}, F_k)$ as: assign a gene level z to class C_l if

$$D_3(\boldsymbol{x}, F_l) = \min_{1<k<K} D_3(\boldsymbol{x}, F_k). \qquad (3)$$

For latter reference, the classification rule in (3) is denoted by D$_3$. The empirical version of $D_3(\boldsymbol{x}, F_k)$ is defined as

$$D_3(\boldsymbol{x}, F_{n_k}) = \sum_{j=1}^{p} \left|\theta_{kj} - x_j\right|, \quad k = 1, 2, \dots, K,$$

where $\theta_{kj} = med(X_{kji}), i = 1, 2, \dots, n_k$.

The almost sure convergence of $D_3(\boldsymbol{x}, F_{n_k})$ to $D_3(\boldsymbol{x}, F_{n_k})$ follows directly from the study of Hall *et al.* (2009). An intuitive property of $D_3(\boldsymbol{x}, F_{n_k})$ is its robustness against outlying observations. $D_3(\boldsymbol{x}, F_{n_k})$ is more robust than $D_2(\boldsymbol{x}, F_k)$. The breakdown point of $\boldsymbol{m}_k$ is 0.5 (see Chakraborty and Chaudhuri (2014) for details).

Hall *et al.* (2009) referred to D$_3$ as median based classifier. A similar classification rule to D$_3$ is quantile discriminant analysis (Hennig and Viroli, 2016). Quantile discriminant analysis, denoted by QuanDA, assigns a test observation to the class with shortest L_1 distance to the *a*th quantile. A similar distance function to $D_1(\boldsymbol{x}, F_{n_k})$ using L_1 norm can be constructed as

$$D_4(\boldsymbol{x}, F_{n_k}) = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\sum_{j=1}^{p} |X_{kij} - x_j|\right).$$

The population version is defined as

$$D_4(\boldsymbol{x}, F_k) = E\left[||\boldsymbol{X}_k - \boldsymbol{x}||_1\right].$$

We define the classification rule based on $D_4(\boldsymbol{x}, F_k)$ as: assign a gene level z to class C_l if

$$D_4(\boldsymbol{x}, F_l) = \min_{1<k<K} D_4(\boldsymbol{x}, F_k). \qquad (4)$$

For latter reference, the classification rule in (4) is denoted by D$_4$.

In literature, other distance based methods include k nearest neighbour rule, shrunken centroid regularized discriminant analysis, nearest shrunken centroid classifier, sparse partial least squares discriminant analysis. k nearest neighbour rule (Cover, 1968), denoted by kNN, assigns each test observation to the class for which the observation have highest representatives among k nearest neighbours. Nearest shrunken centroid classifier (Tibshirani *et al*., 2002), denoted by NSC, assigns each test observation to the class for which the observation achieves the least distance to the class shrunken centroid. Shrunken centroid regularized discriminant analysis (Guo *et al*., 2007), denoted by SCRDA, is very similar to NSC, except for the use of class shrunken regularized centroid in place of class shrunken centroid. SCRDA has been employed in classification of gene expression data. However, SCRDA, NSC and kNN are distance based and are compared with D$_1$, D$_2$, D$_3$ and D$_4$ in our numerical examples, both simulation and analysis of gene expression data.

2. Modified classification rules

The classification rules in (1)-(4) can be modified using a probabilistic approach. The modified method treats a distance measure of any random vector (or gene expression level in micro-array experiment) as a random variable whose distribution function plays a vital role in defining a new classification rule.

Classification rules based on distribution functions of distance functions defined above assign a gene expression level to the class for which it achieves minimum distribution function of some distance measures. Suppose $F_1, F_2, \dots, F_K$ are K distributions of competing classes $C_1, C_2, \dots, C_K$ respectively. Define the distribution function of $U_k^{(1)} = D_1(\boldsymbol{X}_k, F_k)$ as $H\left(\lambda_k^{(1)}\right) = P(U_k^{(1)} \leq \lambda_k^{(1)})$, where $\lambda_k^{(1)} = D_1(\boldsymbol{z}, F_k)$ for $\boldsymbol{z} \in \mathbb{R}^p$. It follows from the definition of $H\left(\lambda_k^{(1)}\right)$ that a central gene expression level will have $H\left(\lambda_k^{(1)}\right)$ close to zero and an extreme gene expression level will have the value of $H\left(\lambda_k^{(1)}\right)$ tends to 1. The classification rule is to assign gene expression level $\boldsymbol{z}$ to class C_l if $H\left(\lambda_l^{(1)}\right) = min_{1<k<K}\, H\left(\lambda_k^{(1)}\right)$. This classification rule is denoted by DD$_1$. The sample version of $H\left(\lambda_k^{(1)}\right)$ denoted by $H\left(\lambda_k^{(1)}\right)$ is defined as

$$\widehat{H}\left(\lambda_{n_k}^{(1)}\right) = \frac{1}{n_k}\sum_{i=1}^{n_k} I\{D_1(\boldsymbol{X}_{ki}, F_{n_k}) \leq D_1(\boldsymbol{x}, F_{n_k})\},$$

where *I* is an indicator function.

Similarly, D_4 can be modified in the version of DD_1. Suppose $H\left(\lambda_k^{(4)}\right) = P(U_k^{(4)} \leq \lambda_k^{(4)})$,where $\lambda_k^{(4)} = D_4(\boldsymbol{z}, F_k)$ and $U_k^{(4)} = D_4(\boldsymbol{X}_k, F_k)$. The classification rule based on $\lambda_k^{(4)}$ is to assign gene expression level $\boldsymbol{z}$ to class C_l if $H\left(\lambda_l^{(4)}\right) = min_{1<k<K} H\left(\lambda_k^{(4)}\right)$. This classification rule is denoted by DD_4. The sample version of $H\left(\lambda_k^{(4)}\right)$ denoted by $\widehat{H}\left(\lambda_{n_k}^{(4)}\right)$ is defined as

$$\widehat{H}\left(\lambda_{n_k}^{(4)}\right) = \frac{1}{n_k}\sum_{i=1}^{n_k} I\{D_4\left(\boldsymbol{X}_{ki}, F_{n_k}\right) \leq D_4\left(\boldsymbol{x}, F_{n_k}\right)\}.$$

3. Simulation Studies

Suppose i th observation is in k th class, $\boldsymbol{X}_{ki} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_k = (\mu_{k1}, \ \mu_{k2}, \dots, \mu_{kp})'$ with $\mu_{1j} = 0$ for $1 \leq j \leq p$, $\mu_{2j} = 0.7$ for $1 \leq j \leq 100$ and $\mu_{2j} = 0$ otherwise and $k = 1, 2$. The covariance structure $\boldsymbol{\Sigma}$ consists of 5×5 blocks, each block of dimension 100×100 with (j, j') element $0.6^{|j-j'|}$. This simulation example is considered in Guo *et al.* (2007).

The second simulation example is similar to the one considered in Hall *et al.* (2009). Suppose $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are two random vectors from distributions F_1 and F_2 respectively. Suppose $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are defined as $X_{11} = \eta_{X_{11}} + U_1$, $X_{12} = \eta_{X_{12}} + U_2$, …, $X_{1p} = \eta_{X_{1p}} + U_p$ and $X_{21} = \eta_{X_{21}} + U_1$, $X_{22} = \eta_{X_{22}} + U_2$, …, $X_{2p} = \eta_{X_{2p}} + U_p$, where U_j, $j = 1, 2, \dots, p$ is student's t distributed with 3 degrees of freedom. We assume $U_1, U_2, \dots, U_p$ are independent. Take $\eta_{X_{11}} = \eta_{X_{12}} = \cdots = \eta_{X_{1p}} = 0$ while $(\eta_{X_{21}}, \eta_{X_{22}}, \cdots, \ \eta_{X_{2p}})$ has first ω non-zero components. Let $\omega = p/4$ such that ω non-zero component η_{Y_k} equals the variance of U_k.

The third simulation example is similar to the first simulation above but considers four classes. The example is described thus: suppose there are four classes π_k, $k = 1, 2, 3, 4$ from $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ such that $\boldsymbol{\mu}_k = (\mu_{k1}, \ \mu_{k2}, \dots, \mu_{kp})'$ with $\mu_{1j} = 0$ for $1 \leq j \leq p$, $\mu_{2j} = 0.7$ for $1 \leq j \leq 100$ and $\mu_{2j} = 0$ otherwise, $\mu_{3j} = 0$ if $1 \leq j \leq 100$, and $\mu_{3j} = 0.5$ if $101 \leq j \leq 200$ and $\mu_{3j} = 0$ otherwise. $\mu_{4j} = 0.5$ if $1 \leq j \leq 100$, and $\mu_{4j} = 0$ if $101 \leq j \leq 300$ and $\mu_{4j} = 0.5$ otherwise. The covariance structure $\boldsymbol{\Sigma}$ consists of 5×5 blocks, each block of dimension 100×100 with (j, j') element $0.6^{|j-j'|}$.

In each of the simulated examples, equal samples are assumed for each of the competing classes. Each experiment consists of measurements on p (=500) features with 50 observations belonging to each training set and 50 observations to each test set. The simulation size is taken to be 500. Mean and standard deviation of proportion of correct classification were computed for each of the methods.

We compare the performance of D_1, D_2, D_3, D_4, DD_1 and DD_4 with k nearest neighbour rule, nearest shrunken centroid classifier (Tibshirani *et al.*, 2002), quantile discriminant analysis (Hennig and Viroli, 2016), shrunken centroid regularized discriminant analysis (Guo *et al.*, 2007) and sparse partial least squares discriminant analysis (splsda) (Chung and Keles, 2010). In our comparison, we choose $k = 1$ and $k = 5$ for implementing kNN using R package *class*. R package *quantileDA* is employed for implementing quantile discriminant analysis. The threshold value for NSC as implemented in R package *pamr* is taken to be 0.5. In these simulation examples, we take the values of parameters of SCRDA to be 0.1 and 0.5 for α and Δ respectively.

The aim of Example 1 is to mimic covariance structure of gene expression data as shown in Guo *et al.* (2007) (see p. 98). Hall *et al.* (2009) argued that gene expression data are naturally tailed and not normally distributed. This forms the basis for Example 2. Table 1 presents the mean and standard deviation of proportions of correct classification. In simulation 1, NSC and D_2 perform best while SCRDA performs worst. D_1, D_3 and D_4 and quanDA perform equivalently. A slight better performance of D_1, D_2 and NSC over D_3 and D_4 may be attributed to the fact that competing classes are multivariate normally distributed and their density functions involve L_2 distance and not L_1 distance.

In simulation 2, D_3 performs best while DD_1 and DD_3 perform worst. The nice performance of D_3 in this example can be attributed to its optimal behaviour when data are tailed. Hall *et al.* (2009) presented theoretical results for the optimal behaviours of D_3 under suitable conditions. QuanDA and D_4 perform equivalently. Among the probabilistic based distance methods, DD_4 competes well. In both simulation examples, $k = 5$ in k nearest neighbour rule performs much better than $k = 1$. It is observed that the performance of D_1-D_4 are much better than the nearest neighbour rules, taken $k = 1$ or $k = 5$. Hall and Pham (2010) argued that for nearest neighbour rules to perform competitively with centroid based classification method (e.g. D_2), dimension must increase slowly as training sample sizes diverge. In simulation

3, the performance of classifiers D_1-D_4 are competitive. SVM and NSC compete well while SCRDA, splsda and 1NN achieve the least proportions of correct classification. It can be inferred from the simulation examples that any of D_1 and D_2 should be used if the distributions of the competing classes are symmetric, and any of D_3 and D_4 if the distributions of the competing classes are tailed.

4. Analysis of Real Data

In this paper, five real life data are used for implementation of the above methods. The datasets are colon tissue data, leukaemia microarray data, small round blue cell tumor data, lymphoma data and prostate cancer data. The datasets are available in R packages *spls*, *rda* and *plsgenomics*.

Colon tissue data set (Alon *et al*., 19999), denoted by colon, contains 62 samples with 2000 genes from two classes. The classes are tumour tissues of size 40 and normal tissues of size 22. The data is split into two equal samples for each class, which constitute the training and test sample for each class. Studies have shown that non-contributing features tend to lower the probability of correctly assigning high dimensional test observations. Shrunken centroid regularised discriminant analysis (SCRDA) (Guo *et al*., 2007) was performed on all samples to remove the non-contributing genes.

Leukaemia microarray data set, denoted by leukaemia, arose from the study of Golub *et al*. (1999) on gene expression levels of 3051 genes for 38 leukaemia patients. The data consists of two groups: group 1 of size 27 and group 2 of size 11. Random training samples of sizes 15 and 7 are selected from groups 1 and 2 respectively. Test samples for the two groups are taken to be complement of the training samples.

Small round blue cell tumor data set, denoted by SRBCT, consists of gene expression level on 2308 genes for 83 patients. This dataset arose from the study of Khan *et al*. (2001) on childhood cancer and is available on R package *rda*. The data set contains four classes; Ewing sarcoma (ES) of size 29, Burkitt lymphoma (BL) of size 11, neuroblastoma (NB) of size 18 and rhabdomyosarcoma (RMS) of size 25. Random training samples of sizes 15, 7, 9 and 15 are taken from classes ES, BL, NB and RMS respectively. Test samples for the completing classes are taken to be complement of the training samples.

The lymphoma dataset (Alizadeh *et al*., 2000) consists of three classes with 4026 genes. The classes are diffuse large B-cell lymphoma (DLBCL) of size 42, follicular lymphoma (FL) of size 9, and chronic lymphocytic leukemia (CLL) of size 11. The lymphoma gene expression data were normalized, imputed, log transformed, and standardized to zero mean and unit variance across genes. Chung and Keles (2010) presented detailed description of the dataset. Random training samples of sizes 30, 6 and 7; and random test samples of sizes 12, 3 and 4 are selected from classes DLBCL, FL and CLL respectively.

Prostate cancer data (Singh *et al*., 2002) consists of two classes (normal and tumor) of sizes 50 and 52 with 6033 genes. The gene expression data and arrays were normalized, log transformed, and standardized to zero mean and unit variance across genes as discussed in Chung and Keles (2010). A random training sample of size 30 is selected from each of the two groups. Random test samples of sizes 20 and 22 are selected from class 1 and 2 respectively.

For each of the datasets, proportion of correct classification was computed for each of the classification methods. The experiment was repeated 1000 times. Mean and standard deviation of proportion of correct classification were computed. The R codes for the competing classification methods are freely available at https://github.com/osMakinde/gene_classify. Table 2 presents the mean and standard deviation of proportions of correct classification of some gene expression data. For leukaemia data, D_1, D_2, D_3, D_4 and NSC perform best while other classifiers perform competitively.

For colon data, splsda and SCRDA perform least while D_1, D_2, D_3, D_4 and quanDA perform equivalently. Other classifiers perform competitively for colon cancer data. For small round blue cell tumor data, splsda and quanDA do not perform well. The near perfect classification is observed in D_1, D_2, D_3, D_4 and NSC for SRBCT data. For lymphoma data, NSC, D_2 and D_3 achieve perfect classification because the three classifiers assign all the test gene expression levels in all repetitions correctly. Other competing classifiers perform well in all repetition. For prostate cancer data, NSC, splsda, D_1, D_2, D_3 and D_4 perform better than other classification methods.

It is observed that D_3 demonstrates a better performance in terms of proportion of correct classification over quanDA as shown in Table 2. This signifies that the use of componentwise median in gene expression classification yields a better performance in distance-based classification than quantile based counterpart. In all the real data examples, D_1, D_2, D_3 and D_4 perform better than DD_1 and

DD_4. However, DD_1 and DD_4 performs competitively with splsda and quanDA.

One of the advantages of distance based classifiers (D_1, D_2, D_3 and D_4) over NSC, QuanDA, splsda, SCRDA and SVM is its computational simplicity and cost in terms of computation time. Figure 1 presents computation time for implementing various classification methods for gene expression data over 1000 repetitions. It is shown in Figure 3 that D_1, D_2, D_3 and D_4 are least computationally expensive. DD_1 and DD_4 generate smaller computational cost than any of NSC, QuanDA, splsda, SCRDA and SVM but higher computation time than D_1, D_2, D_3 and D_4. For distance to mean based method (D_2), we observe the least computation time for all the datasets except Lymphoma data. QuanDA achieves the highest computation time for all the datasets. splsda and SVM also have high computation time. Distance based classification rule based on sample mean vector may be affected by outlying gene expression levels in the data. This has been discussed in literature. However, the difficulty can be overcome using trimmed mean (see Hubert and Van Driessen (2004) for parametric approach and Masse (2009) for nonparametric approach).

Table 1: Mean and standard deviation, in parenthesis of proportions, of correct classification of simulated data examples.

Example	SVM	splsda	QuanDA	SCRDA	INN	5NN	NSC	D_1	D_2	D_3	D_4	DD_1	DD_4
1	0.845	0.8037	0.8071	0.5252	0.6669	0.7328	0.865	0.8383	0.853	0.8228	0.8287	0.7776	0.7737
	(0.04)	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
2	0.8962	0.7583	0.9098	0.7429	0.6727	0.7246	0.8875	0.6632	0.8717	0.966	0.9229	0.6532	0.83
	(0.04)	(0.08)	(0.04)	(0.07)	(0.06)	(0.07)	(0.05)	(0.16)	(0.05)	(0.02)	(0.09)	(0.08)	(0.08)
3	0.8116	0.5785	0.7779	0.4553	0.5532	0.6374	0.8136	0.8025	0.8329	0.801	0.8068	0.7415	0.7349
	(0.03)	(0.05)	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)

Table 2: Mean and standard deviation, in parenthesis of proportions, of correct classification of real data examples.

Dataset	SVM	splsda	quanDA	SCRDA	NSC	D_1	D_2	D_3	D_4	DD_1	DD_4
Leukaemia	0.9995	0.9491	0.9764	0.9894	0.9985	0.9977	0.9991	0.9949	0.998	0.9206	0.9275
	(0.01)	(0.05)	(0.04)	(0.03)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.08)	(0.08)
Colon	0.9138	0.8366	0.9171	0.8419	0.8955	0.9111	0.9161	0.9251	0.9234	0.891	0.8958
	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.06)	(0.05)
SRBCT	0.6574	0.6705	0.6899	0.9778	0.9991	0.9236	0.9746	0.9941	0.9802	0.8466	0.8878
	(0.06)	(0.10)	(0.06)	(0.03)	(0.01)	(0.07)	(0.03)	(0.02)	(0.04)	(0.07)	(0.06)
Lymphoma	0.9879	0.95	0.9887	0.9999	1.0000	0.9998	1.0000	1.0000	0.9999	0.9494	0.9517
	(0.03)	(0.05)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)	(0.05)
Prostate	0.94	0.9368	0.8596	0.8902	0.9452	0.934	0.9486	0.9308	0.931	0.8801	0.8863
	(0.03)	(0.03)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.06)	(0.05)

5. Conclusion

Some distance functions are surveyed, formulated and discussed in this paper with aim of presenting computationally simple and efficient classification methods for gene expression data. The distance functions employed in the classification methods include average distance and its L_1 version, L_2 distance to class mean vector, L_1 distance to class componentwise median and their distribution based versions. All these classifiers are computationally simple and very cheap in terms of computational cost and their results can be easily interpreted. The distance based methods discussed in this paper perform competitively with some other classification methods in literature as illustrated in our numerical examples. This supports a claim (Hand, 2006) that simple classification methods like D_1, D_2, D_3, D_4, DD_1 and DD_4 tend to have comparative performance to more complicated classification methods. Also, the proposed methods can be implemented for multiclass extensions.

In order to implement the classification methods based on distribution function of some distance measures discussed in this paper, distance of each training observation in each class is first computed with respect to other observations in the class. Second, distance of each test observation is computed with respect to training observations in each competing class. Third, compute the distribution function of distance of each test observation. The test observations are then assigned to the class with the least distribution value.

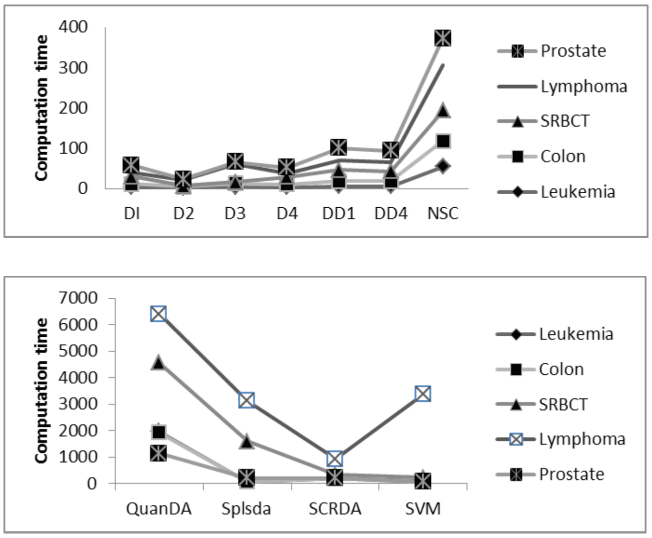

Fig. 1. Computation time (in seconds) for implementing various classification methods and computing proportions of correct classification of the methods for gene expression data over 1000 repetitions.

References

Alizadeh, A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H. … Staudt, L.M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, **403**(6769):503-511

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. & Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America, **96**(12):6745-6750

Chakraborty, A. & Chaudhuri, P. (2014). The deepest point for distributions in infinite dimensional spaces. Statistical Methodology, **20**:27-39

Chung, D. & Keles, S. (2010). Sparse Partial Least Squares Classification for High Dimensional Data. Statistical Applications in Genetics & Molecular Biology, **9**(1), Article 17.

Colak, C., Colak, M.C., Ermis, N., Erdil, N. & Ozdemir, R. (2016). Prediction of cholesterol level in patients with myocardial infarction based on medical data mining methods. Kuwait Journal of Science, **43**(3):86-90.

Cover, T.M. (1968). Rates of convergence for nearest neighbor procedures. Proc. Hawaii Int'l Conf. Systems Sciences. Western Periodicals, Honolulu. 413-415.

Dabney, A.R. (2005). Classification of microarrays to nearest centroids. Bioinformatics, **21**(22):4148-4154.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D. (2000). Support Vector Machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, **16**:906-914.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, **286**:531-537.

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

Hall, P., Titterington, D.M. & Xue, J. (2009). Median

Based classifiers for High Dimensional Data. Journal of the American Statistical Association, **104**(488):1597-1608.

Hall, P. & Pham, T. (2010). Optimal properties of centroid-based classifiers for very high-dimensional data. The Annals of Statistics, **38**(2):1071-1093.

Hand, D.J. (2006). Classifier technology & the illusion of progress, Statistical Science, **21**(1):1-14.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). The elements of statistical learning: data mining, inference and prediction. Springer, New York. Chapter 16.

Hennig, C. & Viroli, C. (2016). Quantile-based classifiers. Biometrika, **103**(2):435-446.

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

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. & Meltzer, P.S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, **7**:673-679.

Klassen, M. & Kim, N. (2009). Nearest shrunken centroid as feature selection in microarray data. Proceeding of computers and their applications, 227-232.

Li, B. and Yu, Q. (2008). Classification of functional data: A segmentation approach. Computational Statistics and Data Analysis, **52**:4790-480

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

Masse, J.C. (2009). Multivariate Trimmed means based on the Tukey depth, Journal of Statistical Planning and Inference, **139**(2):366-384

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., Damico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. & Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behaviour. Cancer Cell, **1**:203-209.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002). Diagnosis of multiple cancer type by shrunken centroid. Proceedings of the National Academy of Sciences, USA, **99**(10):6567-6572.

Vanitha, C.D.A., Devaraj, D. & Venkatesulu, M., (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Computer Science, **47**:13-21.

Wang, L., Zhu, J. & Zou, H. (2008). Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics, **24**(3):412-419.

Submitted: 03/01/2018
Revised: 13/05/2018
Accepted: 27/11/2018

تصنيف بيانات التعبير الجيني: بعض الطرق القائمة على المسافة

أولوسولا صموئيل ماكيندي
قسم الاحصاء، الجامعة الفيدرالية للتكنولوجيا، أكور، نيجيريا
المؤلف: osmakinde@futa.edu.ng

ملخص

مجموعة بيانات micro-array هي مثال كلاسيكي لبيانات الإنتاجية العالية التي تتميز بمزيد من الخصائص (الجينات) أكثر من نقاط العينة (مستويات التعبير الجيني). تم اقتراح عدد من تقنيات التصنيف في النشرات العلمية، وكانت العديد من هذه الطرق إما مكلفة حسابياً أو كان أداؤها دون المستوى الأمثل. في هذا البحث، تم النظر في بعض دوال المسافة وتمت صياغة قواعد التصنيف على أساس دوال المسافة؛ والتي تشمل: قياس متوسط المسافة، وسيط المسافة إلى العنصر، والمتوسط إلى المسافة. وتم كذلك تحديد نهج احتمال لقواعد التصنيف على أساس اثنين من قياسات المسافة. وتم استخدام تقنية اختيار الجينات التي تستند إلى تحليل مميز منظم مركزي منقبض على أنسجة خلايا زرقاء صغيرة مستديرة وسرطان القولون وسرطان الغدد الليمفاوية وسرطان البروستاتا وسرطان الدم قبل تطبيق قواعد التصنيف. وأُجريت ثلاث دراسات محاكاة لتقليد بيانات التعبير الجيني. وتمت مقارنة أداء طرق التصنيف المذكورة أعلاه مع أداء بعض طرق التصنيف المعروفة في النشرات العلمية. وكان أداء طرق التصنيف عن بعد منافساً لبعض طرق التصنيف الحالية. وكانت الطرق المستندة على المسافة التي تم تنفيذها في هذه الدراسة بسيطة من الناحية الحسابية ورخيصة جداً من حيث التكلفة.