Scholars Journal of Physics, Mathematics and Statistics **ISSN 2393-8056 (Print)**
Sch. J. Phys. Math. Stat. 2017; 4(4):205-210 **ISSN 2393-8064 (Online)**

(An International Publisher for Academic and Scientific Resources)

A New Dimension Reduction Approach Based on Distance for Mixture Discriminant Analysis of the High-Dimensional Data

Ulku Erisoglu, Aydın Karakoca*, Ahmet Pekgor and Murat Erisoglu
Necmettin Erbakan University, Faculty of Science, Department of Statistics, Konya, Turkey

***Corresponding author**
Aydın KARAKOCA

Article History
Received: 24.11.2017
Accepted: 05.12.2017
Published: 30.12.2017

DOI:
10.21276/sjpms.2017.4.4.9

Abstract: In this study, we proposed a novel dimension reduction approach for mixture discriminant analysis on based mixture of multivariate normal distributions of high-dimensional data. We considered case of a classification problem that the number of observations (n) is less than the number of variables (p). The proposed approaches compared with classical dimension reduction methods such as F approach, principal component analysis, clustering of variables and multidimensional scaling.
Keywords: Dimension Reduction, High-dimensional data, Mixture discriminant analysis, F approach, Clustering of variables, SMACOF

INTRODUCTION

A classic problem in data analysis is classifying high dimensional data into multiple predefined categories. In particular, two-group classification problem has received a great deal of attention. Many methods have been proposed. But in various situations, especially when analyzing data with a small sample size relative to the number of variables (e.g. medical image, genetic microarray, chemometrics and text classification), many classification techniques become impractical. For example, Fisher's discriminant analysis is not applicable if the number of input variables is greater than the number of observations. Other methods, even some sophisticated methods, such as neural networks and support vector machines, do not explicitly require the data dimension smaller than the sample size, but give poor classification accuracy in practice when the data dimension is ultra-high, as in fMRI and microarray data. Moreover, for the sake of model simplicity, a concise relationship between input variables and the response is required to achieve a better model interpretation [10].

A natural way to deal with high dimensional classification problems is to first reduce the data to a lower dimensional subspace and then apply some standard classification strategy, such as linear discriminant analysis or logistic regression, to the reduced data.

Many methods use global dimension reduction techniques to overcome problems due to high dimensionality. A widely used solution is to reduce the dimension of data before using a classical classification method [3]. Dimension reduction techniques can be divided into techniques for variable (feature) extraction and variable selection.

Variable extraction techniques build new variables carrying a large part of the global information. Variable transformation techniques attempt to summarize a dataset in fewer dimensions by creating combinations of the original attributes. These techniques are very successful in uncovering latent structure in datasets. Among these techniques, the most popular one is principal component analysis (PCA).

Variable selection plays an important role in classification [6]. Before beginning designing a classification method, when many variables are involved, only those variables that are really required should be selected; that is, the first step is to eliminate the less significant variables from the analysis. Various variable selection schemes have been applied to high dimensional data with a double purpose [2]. Variable selection may be performed as a preliminary step before classification, because the chosen classification method works only with a small subset of variables. Variable selection is of crucial interest for researchers who want to identify significant variables which are associated with studies.

In this study, we proposed a novel dimension reduction approach for mixture discriminant analysis on based mixture of multivariate normal distributions of high-dimensional data. The proposed approaches compared with classical dimension reduction methods such as *F* approach, principal component analysis, clustering of variables and multidimensional scaling.

Notations

$X_1, \dots, X_p$ denote the variables. In this study, they are continuous variables. $\mathbf{x} = (X_1, \dots, X_p)$ denotes the corresponding random vector. Y denotes the class membership. $(\mathbf{x}_i, Y_i)_{i=1,\dots,n}$ is observed data set, with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ denoting measurements of the p variables and Y_i the class membership for observation *i*-th.

For $k = 1, \dots, K$, $\mathbf{x}_{k1}, \dots., \mathbf{x}_{kn_k}$ denote the observations from class k, where n_k is the number of observations from class k and $k1, \dots, kn_k$ are the indices of the observations from class in the data set $(\mathbf{x}_i, Y_i)_{i=1,\dots,n}$. Thus, for $k = 1, \dots, K$, and $i = 1, \dots, n_k$ one has $Y_k = k$. $\mathbf{X}$ is the $n \times p$ matrix which contains $\mathbf{x}_i$ in its *i*-th row, for $i = 1, \dots, n$. In the following

$$\boldsymbol{\mu} = \mathrm{E}(\mathbf{x}) = (\mu_1, \dots, \mu_p) \quad (1)$$

denotes the mean vector of $\mathbf{x}$ and $\hat{\boldsymbol{\mu}} = \frac{1}{n}\mathbf{I}_n\mathbf{X} = (\hat{\mu}_1, \dots, \hat{\mu}_p)$ where $\mathbf{I}_n$ is the vector of ones of length n. Σ is the $p \times p$ covariance matrix of $\mathbf{X}$,

$$\Sigma = \mathrm{cov}(\mathbf{x}) = \mathrm{E}\big((\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})\big) \quad (2).$$

$\mathbf{S}$ denotes the unbiased estimator of Σ;

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) \quad (3).$$

Dimensionality Reduction Techniques

Dimensionality reduction is the process of finding a suitable lower dimensional space in which to represent the original data. The linear dimensionality reduction is performed using a linear transformation of the form:

$$\mathbf{Z} = \mathbf{XA} \quad (4)$$

where $\mathbf{A}$ is an $p \times d$ transformation matrix which projects $\mathbf{X}$ onto a d dimensional subspace $\mathbf{Z}$ ($d < p$). However, there are many possible approaches to choose $\mathbf{A}$.

Probably the most commonly applied method in the category is principal component analysis (PCA). In the PCA, $\mathbf{A}$ is the transformation matrix which contains d eigenvectors corresponding to d highest eigenvalues of the covariance matrix of $\mathbf{x}$ [11]. The main purpose of PCA is to reduce the dimensionality from p to d where $d < p$, while at the same time accounting for as much of the variation in the original data set as possible. With PCA, we transform the data to a new set of coordinates or variables that are a linear combination of the original variables. In addition, the observations in the new principal component space are uncorrelated.

Multidimensional scaling (MDS) is another approach used to dimensionality reduction. In general, MDS is a set of techniques for the analysis of proximity data measured on a set of objects in order to reveal hidden structure. The purpose of MDS is to find a configuration of the data points in a low-dimensional space such that the proximity between objects in the full-dimensional space is represented with some degree of fidelity by the distances between points in the low-dimensional space. This means that observations that are close together in a high-dimensional space should be close in the low-dimensional space [8].

There are a lot of different algorithms to solve MDS problem. Scaling by MAjorizing a COmplicated Function (SMACOF) is one of them. SMACOF is an iterative majorization algorithm to solve MDS problem with STRESS criterion. For the mathematical details of SMACOF algorithm, please refer to [1].

The variable selection methods found in the literature can be divided into two distinct groups: univariate ranking methods and optimal subset selection. *F* approach is one of univariate ranking methods[12]. Each variable is taken individually and a relevance score measuring the discriminating power of the variable is computed. The variables are then ranked according to their score. One can choose to select only the d top-ranking variables (where $< p$) or the variables whose score exceeds a given threshold. One of the most common relevance scores is the *F* test statistic. For variable *j*-th, the *F* test statistic is defined as

$$F_j = \frac{(n-k)\sum_{k=1}^{K} n_k\left(\hat{\mu}_{kj} - \hat{\mu}_j\right)^2}{(K-1)\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(x_{ij} - \hat{\mu}_{kj}\right)^2} \quad (5)$$

F_j test statistic is F distributed with degrees of freedom $K-1$ and $n-K$. The corresponding p value can be used as a relevance score.

Clustering of variables (Cluster Approach) is another approach used to dimensionality reduction. In this approach, dimensionality reduction provided with using centers of clustering variables instead of the original variables. There are many several algorithms in cluster analysis. We used k- means clustering algorithm in this study.

The Proposed Algorithm

In this section, the proposed algorithm for dimensionality reduction is explained.

Step 1: Firstly, empirical mean vector of **x** within each class is computed. For $k = 1, \dots, K$, n_k is the number of observations in class k. $\boldsymbol{\mu_k}$ denotes the mean vector of within class k.

$$\boldsymbol{\mu}_k = \mathrm{E}(\mathbf{x}|\mathrm{Y} = k) = \left(\mu_{k1}, \dots, \mu_{kp}\right) \quad (6)$$

and $\hat{\boldsymbol{\mu}}_k = \left(\hat{\mu}_{k1}, \dots, \hat{\mu}_{kp}\right)$ is the empirical mean vector of **x** within class k .

$$\hat{\mu}_{kj} = \frac{1}{n_k}\sum_{x_{ij}\in k} x_{ij} \quad \text{for } j = 1, \dots, p \quad (7).$$

The computed empirical mean vectors for each class k will used to as the axis in variable selection.

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_{11} & \cdots & \hat{\mu}_{k1} \\ \vdots & \ddots & \vdots \\ \hat{\mu}_{1p} & \cdots & \hat{\mu}_{kp} \end{bmatrix} \quad (8).$$

Step 2: The mean of data points is calculated as the center of the dataset according to selected axis after determining the axis.

$$\mathbf{c} = [c_1 \quad \cdots \quad c_k] \quad (9)$$

$$c_k = \frac{1}{p}\sum_{j=1}^{p} \hat{\mu}_{kj} \quad (10)$$

Step 3: The Euclidean distance between each data point and the center is created by

$$d_{j,\mathbf{c}} = \left\{\left(\hat{\mu}_{j1} - c_1\right)^2 + \cdots + \left(\hat{\mu}_{jk} - c_k\right)^2\right\}^{\frac{1}{2}} \quad \text{for } j = 1, \dots, p \quad (11).$$

A data point $\mathbf{v}_1$ which has the highest distance will be selected as the first candidate variable.

Step 4: The Euclidean distances between each data points and $\mathbf{v}_1$ are calculated by

$$d_{j,\mathbf{v}_1} = \left\{\left(\hat{\mu}_{j1} - v_{11}\right)^2 + \cdots + \left(\hat{\mu}_{jk} - v_{1k}\right)^2\right\}^{\frac{1}{2}} \quad \text{for } j = 1, \dots, p \quad (12).$$

The data point with the highest distance of $d_{j,\mathbf{v}_1}$ will be selected as the second candidate variable $\mathbf{v}_2$.

Step 5: To select a next candidate variable, The Euclidean distances between the rest data points and the mean of selected candidate points are calculated.

$$d_{j,\left(\frac{\mathbf{v}_1+\mathbf{v}_2}{2}\right)} = \left\{\left(\hat{\mu}_{j1} - \frac{v_{11}+v_{21}}{2}\right)^2 + \cdots + \left(\hat{\mu}_{jk} - \frac{v_{1k}+v_{2k}}{2}\right)^2\right\}^{\frac{1}{2}} \quad (13).$$

The data point with the highest distance of $d_{j,\left(\frac{\mathbf{v}_1+\mathbf{v}_2}{2}\right)}$ will be selected as the third candidate variable $\mathbf{v}_3$.

The process is repeated until the number of candidate variables equals to the predefined number of dimensions.

Mixture Discriminant Analysis

In the mixture discriminant analysis, suppose we have observation n_k from population k for $k = 1, \dots, K$. Each class k is divided into R_k artificial subclasses. According to this clustered approach, each subclass has a multivariate normal distribution $\mathbf{x}_i \sim N(\boldsymbol{\mu}_{kr}, \Sigma_{kr})$ with its own mean vector $\boldsymbol{\mu}_{kr}$ and Σ_{kr} is covariance matrix for the r-th subclass in k-th class. The prior probability for class k is π_k and π_{kr} is the mixing probability for the r-th subclass in in k-th class, such that $\sum_{r=1}^{R_k} \pi_{rk} = 1$. Then mixture density for class k is

$$m_k(\mathbf{x}) = \mathbf{P}(\mathbf{X} = \mathbf{x}|Y = k) = |2\pi\Sigma_k|^{-\frac{1}{2}} \sum_{r=1}^{R_k} \pi_{kr} exp\left[-D(\mathbf{x} - \boldsymbol{\mu}_{kr})/2\right] \quad (14)$$

where $D(\mathbf{x} - \boldsymbol{\mu}_{kr}) = (\mathbf{x} - \boldsymbol{\mu}_{kr})\Sigma_{kr}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{kr})^T$ is Mahalanobis distance. The posterior probabilities are obtained, based on Bayes rule, such that

$$P(Y = k|\mathbf{X} = \mathbf{x}) \sim \boldsymbol{\pi_k} \sum_{r=1}^{R_k} \pi_{kr} exp\left[-D(\mathbf{x} - \boldsymbol{\mu}_{kr})/2\right] \quad (15)$$

where π_k is the prior probability for class k. An observation is classified into the class k which has the highest posterior probability. The discrimination rules depend on the unknown parameters which are to be estimated from the data [7].

Application and Conclusions

The datasets and features which will use in the comparison of the proposed algorithm with the investigated dimension reduction methods are given Table 1.

Table-1: The datasets and features

Data sets	n	p	k	n_k
Multi	103	4576	4	26, 26, 28, 23
Apple	60	701	3	20, 20, 20
Cherry	60	701	3	20, 20, 20
Chowdary	104	22283	2	62, 42

The multi dataset [9] contains in total 103 samples in four classes which have 26, 26 28, 23 samples, respectively. Each sample contains 4576 genes. Spectral reflectance data from the wavelength range of 325–1025 nm with 701 spectral features were collected from the apple and cherry trees using a visible-near infrared spectroradiometer [5]. The Chowdary data set is composed by tissue from lymph nodenegative breast tumors and Dukes' B colon tumors [4].

High-dimension datasets are reduced by dimension reduction methods and the reduced data sets are applied the mixture discriminant analysis on based mixture of multivariate normal distributions. After, classification accuracies are computed for the each mixture discriminant analysis. The classification accuracies are given Tables 2-6 according to number of dimension for dimension reduction methods.

Table-2: The classification accuracies according to number of dimension for F approach

Data sets / d	5	6	7	8	9	10	11	12
Multi	0.8350	0.9223	0.9903	0.9903	0.9903	0.9806	0.9903	0.9806
Apple	0.6667	0.6833	0.8000	0.9167	0.9000	0.9333	0.9667	0.9833
Cherry	0.8000	0.8500	0.9167	0.9500	0.9500	0.9667	0.9667	0.9667
Chowdary	0.9615	0.9904	0.9904	0.9904	0.9904	0.9904	0.9904	0.9904

Table-3: The classification accuracies according to number of dimension for clustering of variables

Data sets / d	5	6	7	8	9	10	11	12
Multi	0.9612	0.9903	0.9903	0.9903	1.0000	1.0000	1.0000	1.0000
Apple	0.8167	0.9000	0.9167	0.9333	0.9500	0.9667	0.9833	1.0000
Cherry	0.7333	0.7667	0.8000	0.8833	0.9000	0.9167	0.9667	1.0000
Chowdary	0.7404	0.8750	0.8558	0.8942	0.9135	0.9135	0.9327	0.9519

Table-4: The classification accuracies according to number of dimension for the principal component analysis

Data sets / d	5	6	7	8	9	10	11	12
Multi	0.9709	0.9806	0.9903	1.0000	1.0000	1.0000	1.0000	1.0000
Apple	0.8333	0.9000	0.9500	0.9833	0.9833	0.9667	0.9833	1.0000
Cherry	0.6833	0.7500	0.7833	0.8667	0.8833	0.9667	0.9500	1.0000
Chowdary	0.8654	0.8365	0.8654	0.8942	0.8942	0.9423	0.9519	0.9615

Table-5: The classification accuracies according to number of dimension for the SMACOF (Distance: City-Block)

Data sets / d	5	6	7	8	9	10	11	12
Multi	0.9709	0.9806	0.9806	0.9903	0.9903	1.0000	1.0000	1.0000
Apple	0.8500	0.8667	0.8833	0.8833	0.9000	0.9167	0.8833	0.9167
Cherry	0.7000	0.6833	0.7833	0.7833	0.8833	0.8833	0.9833	0.9833
Chowdary	0.8942	0.8365	0.8654	0.8942	0.8942	0.9423	0.9423	0.9519

Table-6: The classification accuracies according to number of dimension for the proposed algorithm

Data sets / d	5	6	7	8	9	10	11	12
Multi	0.9612	0.9806	0.9903	0.9903	1.0000	1.0000	1.0000	1.0000
Apple	0.7167	0.7833	0.8000	0.8500	0.9167	0.9000	0.9833	0.9833
Cherry	0.7833	0.9000	0.9167	0.9833	1.0000	1.0000	1.0000	1.0000
Chowdary	0.9520	0.9519	0.9615	0.9712	0.9904	0.9904	1.0000	1.0000

The comparison of dimension reduction methods according to classification accuracy in mixture discriminant analysis for of each dataset is given Figure 1.

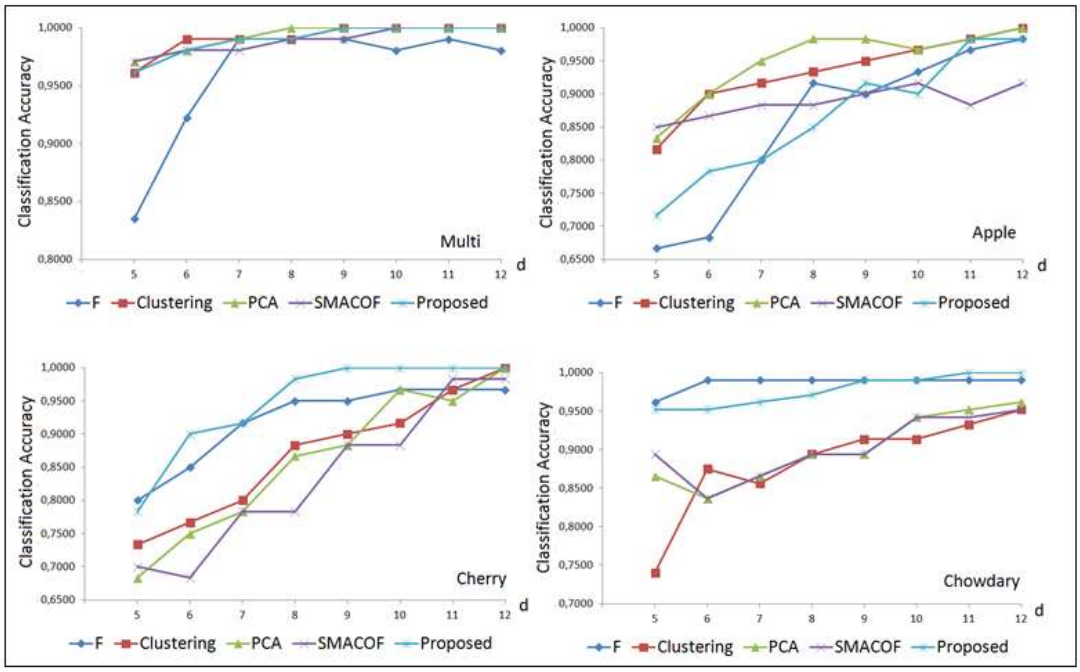

Fig-1: The comparison of dimension reduction techniques according to classification accuracy in mixture discriminant analysis

As result of the study, the proposed dimensional reduction approach is shown good performance for mixture discriminant analysis on based mixture of multivariate normal distributions in terms of criteria classification accuracy. Increasing number of the used dimensional for classification decreases the difference among the performance of methods as expected. It is proved that proposed dimensional reduction approach is useful in dimensional reduction for mixture discriminant analysis of high dimensional datasets.

REFERENCES

1. Borg I, Groenen PJ. Modern multidimensional scaling: Theory and applications. Springer Science & Business Media; 2005 Aug 4.
2. Boulesteix AL. Dimension reduction and classification with high-dimensional microarray data (Doctoral dissertation, lmu). 2005.
3. Bouveyron C, Girard S, Schmid C. High-dimensional data clustering. Computational Statistics & Data Analysis. 2007 Sep 15;52(1):502-19.
4. Chowdary D, Lathrop J, Skelton J, Curtin K, Briggs T, Zhang Y, Yu J, Wang Y, Mazumder A. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. The journal of molecular diagnostics. 2006 Feb 28;8(1):31-9.
5. Dedeoğlu M. Detection of Zinc Deficiency in Apple And Cherry Trees By Visible Near Infrared Spectroradiometry. Ms Thesis. Selçuk University. 2011.
6. Fan J, Fan Y. High dimensional classification using features annealed independence rules. Annals of statistics. 2008;36(6):2605.
7. Hastie T, Tibshirani R. Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society. Series B (Methodological). 1996 Jan 1:155-76.
8. Martinez WL, Martinez AR, Martinez A, Solka J. Exploratory data analysis with MATLAB. CRC Press; 2010 Dec 16.
9. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF, Hampton GM. Molecular classification of human carcinomas by use of gene expression signatures. Cancer research. 2001 Oct 15;61(20):7388-93.
10. Tian TS, Wilcox RR, James GM. Data reduction in classification: A simulated annealing based projection method. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2010 Oct 1;3(5):319-31.
11. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics. 2001 Sep 1;17(9):763-74.
12. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005 Apr 1;67(2):301-20.