Applications and Applied Mathematics: An International Journal (AAM)

Volume 18 | Issue 1 Article 8

6-2023

(R2025) Improving the LDA Linear Discriminant Analysis Method by Eliminating Redundant Variables for the Diagnosis of COVID-19 Patients

Kianoush Fathi Vajargah
Islamic Azad University

Hamid Mottaghi Golshan
Islamic Azad University

Fazel Badakhshan Farahabadi
Islamic Azad University

Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466

Applications and Applied Mathematics:
An International Journal
(AAM)

Vol. 18, Issue 1 (June 2023), Article 8, 10 pages

Improving the LDA Linear Discriminant Analysis Method By Eliminating Redundant Variables for the Diagnosis Of COVID-19 Patients

[1,*]Kianoush Fathi Vajargah, [2]Hamid Mottaghi Golshan, and [3]Fazel Badakhshan Farahabadi

[1]Department of Statistics
Islamic Azad University
Science and Research Branch
Tehran, Iran
K_fathi@iau-tnb.ac.ir

[2]Department of Mathematics
Shahriar Branch
Islamic Azad University
Shahriar, Iran
Ha.Mottaghi@iau.ac.ir

[3]Department of Statistics
Islamic Azad University
Tehran North Branch
fazelbadakhshan@gmail.com

*Corresponding Author

Received: August 26, 2022; Accepted: February 2, 2023

Abstract

Nowadays, with the increase in data production speed, the process of data analysis has faced many problems because this big data is often accompanied by plug-in data and redundant data. Therefore, the use of dimensional methods in the pre-data analysis stage is necessary. In data mining, dimensional reduction is one of the most important steps in data pre-processing. Principal component analysis (PCA) and linear discriminant analysis (LDA) are often used to reduce dimensions in data mining. The LDA method is a monitored and controlled method but the PCA is not controlled method. When the number of samples in classes is large and when training data is uniformly distributed, LDA works better. LDA is a traditional statistical method for classification. In this research, we improve the LDA method by identifying and removing redundant variables, Then we use this improved LDA method to classify the diagnosis data of Covid-19 patients.

Keywords: Linear discriminant analysis (LDA); Gaussian copula; Naïve Beyes; KNN; Covid-19

MSC 2010 No.: 62H05, 62H25, 62H20

1. Introduction

In this study, we used the LDA method to reduce dimensions and classification of the data, and we tried to improve the LDA method by using a method to identify and remove redundant variables. Research has been done to improve the LDA method (Zheng et al. (2020); Lei et al. (2019)). In this study, we use the copula function to identify redundant variables so that we can measure the structural dependence and nonlinear dependencies of the variables. And we remove variables that are highly structurally dependence to each other, as redundant or noise variables. Because they behave similarly, then we use the LDA method for the remaining variables and classify the data.

To determine the structural dependence of variables, we use the copula function estimation, so that by fitting the appropriate copula function and estimating the copula function parameter, we determine the structural correlation of the variables, And then we put the structural depended variables in a subset (Badakhshan Farahabadi et al. (2021); Saoudi et al. (2016)). According to the appropriate criteria, we select the appropriate variable from this subset, and we remove the other variables in this subset as redundant variables and noise variables.

Finally, we use the LDA method to classify data related to the rapid detection of Covid-19 by cough frequency, which can be used in smart home devices. We also compare it with conventional classification methods in data mining such as Naïve Bayes and KNN, to see that this two-step method can improve the accuracy of classification.

2. Linear discriminant analysis (LDA)

Linear discriminant analysis is a statistical method for reducing the dimensions of a problem and classifying clusters by maximizing the scatters between groups to scatters within groups. Linear discriminant analysis is in fact similar to and borrowed from the method used by Ronald Fisher (Fisher (1936)) to determine the degree of differentiation between groups and analysis of variance.

The purpose of linear discriminant analysis is to find a projection or transformation w. On the data set A so that it can maximize the scatters between groups to the scatters within groups the converted data (Rao (1948)). In this case, if we consider the matrix A to be the set of initial d-dimension data, the purpose of performing linear discriminant analysis is to find a vector or matrix w that can maximize the ratio for the converted data Y (Li and Schonfeld (2014)),

$$Y = w^T A.$$

2.1. Definitions related to linear discriminant analysis

(1) Scatters Between Groups:
Assume that the value for the observations is denoted by x_j and the center of group c is denoted

by μ_c. Then the method of calculation S_w is in accordance with the following,

$$S_w = \sum_c \sum_{x_j \in c} (x_j - \mu_c)(x_j - \mu_c)^T.$$

(2) Scatters Within Groups:
We use the following equation to calculate the scatters between groups,

$$S_B = \sum_{c \in \text{classes}} N_c(\mu_c - \mu)(\mu_c - \mu)^T,$$

where N_c is the number of members of each group or category and the total average is μ.

In the LDA method, we look for a linear combination of observations based on which we can maximize the ratio $\frac{S_B}{S_w}$ for new and converted data. Of course, it is clear that the transformed data has a smaller dimension than the original data. In this way, the dimensions of the problem are reduced and we can make a better diagnosis for the categories or groups by considering the smaller dimensions of the data that are formed in the following relation,

$$y = w^T x.$$

Since this conversion (multiplication) by the vector or matrix w affects all points, it also affects the group average and the total average, which are in the form μ and μ_c.

By applying the conversion matrix w, we calculate the scatters within the groups as follows,

$$S_w = \sum \sum (w^T(x_j - \mu_c))(w^T(x_j - \mu_c))^T = w^T S_w w.$$

And scatters between groups is also calculated as follows,

$$S_B = \sum_{c \in \text{classes}} N_c(w^T(\mu_c - \mu))(w^T(\mu_c - \mu))^T = w^T S_B w.$$

Thus, the ratio of scatters between groups to within groups is calculated as follows,

$$\frac{w^T S_B w}{w^T S_w w}.$$

To maximize this ratio, we can maximize the numerator and consider the denominator as constant ($w^T S_w w = k$) and its Lagrangian form with parameter λ is as follows,

$$L = w^T S_B w - \lambda(w^T S_B w - k).$$

After derivation we have,

$$\begin{aligned} &\frac{\partial L}{\partial w} = S_B w - \lambda S_w w = 0, \\ &S_w^{-1} S_B w = \lambda w, \\ &(S_w^{-1} S_B - \lambda I) w = 0. \end{aligned}$$

If there is an inverse matrix S_w the answers to this equation are the eigenvalue (λ) and the eigenvectors w for the matrix $S_w^{-1} S_B$ (Farag and Elhabian (2008)).

3. Modeling with copula function

The important application of a copula is to present an appropriate method for generating distributions of correlated random multivariate variables and offer a solution to the problem of density estimation conversion.

The scalar field theory indicates that there is a unique m-dimensional copula in $[0,1]^m$ with standard normal marginal distributions $U_1, \ldots, U_m$ (Durante et al. (2013)), whereas Nelson stated that every distribution function F with margins $F_1, \ldots, F_m$ could be written as follows:

$$\forall (X_1, \ldots, X_m) \in \mathbb{R}^m, \qquad F(X_1, \ldots, X_m) = c(F_1(X_1), \ldots, F_m(X_m)).$$

To evaluate a copula selected with an estimated parameter and avoid defining any hypotheses on $F_i(X_i)$, the empirical distribution function of a marginal distribution $F_i(X_i)$ can be employed to transform m samples of X into m samples of U (Nelsen (2007)).

3.1. Gaussian Copula

The difference between a Gaussian copula and a joint normal distribution is that the Gaussian copula allows us to have different types of a distribution function for a joint distribution. However, according to the probability theory, the multivariate normal distribution is the generalization of a one-dimensional normal distribution.

The standard multivariate Gaussian copula is defined as below:

$$c(\Phi(X_1), \ldots, \Phi(X_m)) = \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} X^T (\Sigma^{-1} - I) X\right),$$

where $\Phi(x_i)$ is the standard distribution of $f_i(x_i)$, whereas $X_i \sim N(0,1)$ and Σ are the correlation matrices. As a result, $c(u_1, \ldots, u_m)$ is called the Gaussian copula, and the joint density is obtained from the following equation:

$$c(u_1, \ldots, u_m) = \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \xi^T (\Sigma^{-1} - I) \xi\right],$$

where $u_i = \Phi(x_i)$ and $\xi = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m))^T$ ((Lopez-Paz et al. (2013); MacKenzie and Spears (2014))).

3.2. Maximum Likelihood Estimation

Consider $Y = (Y_1, \ldots, Y_m)$ a random diagram. Assume that $F_{Y_1}(\cdot|\theta_1), \ldots, F_{Y_m}(\cdot|\theta_m)$ is a parametric model for marginal distribution functions and that $c_Y(\cdot|\theta_C)$ is a parametric model for copula Y. The following equation is true:

$$f_Y(y) = f_Y(y_1, \ldots, y_m) = c_Y(F_{Y_1}(y_1), \ldots, F_{Y_m}(y_m)) \prod_{j=1}^{m} f_{Y_j}(y_j).$$

Assume that an instance of IID is $Y_{1:n} = (Y_1, \ldots, Y_n)$. The likelihood logarithm is then obtained

$$\begin{aligned}\log L(\theta_1, \ldots, \theta_m, \theta_C) &= \log \prod_{i=1}^{n} f_Y(y_i) \\ &= \sum_{i=1}^{n} (\log[c_Y(F_{Y_1}(y_{i,1}|\theta_1), \ldots, F_{Y_m}(y_{i,m}|\theta_m)|\theta_C)] \\ &\quad + \log(f_{Y_1}(y_{i,1}|\theta_1)) + \cdots + \log(f_{Y_m}(y_{i,m}|\theta_m))).\end{aligned}$$

ML estimators $\widehat{\theta_1}, \ldots, \widehat{\theta_2}, \widehat{\theta_C}$ are obtained from the maximization of the above equation based on $\theta_1, \ldots, \theta_m, \theta_C$.

This method has a few setbacks:

(1) There are too many parameters to estimate, especially for large values of m. As a result, optimization can be difficult.
(2) If any of the univariate parametric distributions $F_{Y_i}(\cdot|\theta_i)$ are defined incorrectly, bias can emerge in univariate distributions and the copula (Lei et al. (2019)).

3.3. Pseudo-MLE

Pseudo-MLE helps solve the above-mentioned MLE problems. This method has the following setbacks:

(1) The marginal distribution functions are first estimated to define $\widehat{F}_{Y_j}$, for $j = 1, \ldots, m$. For this purpose, the following two methods can be adopted:
   - The empirical distribution function is defined as below for $y_{1,i}, \ldots, y_{n,j}$:

$$\widehat{F}_{Y_i}(y) = \frac{\sum_{i=1}^{n} I_{\{y_{i,j} \le y\}}}{n+1}.$$

   - A parametric model is developed with $\widehat{\theta_j}$ obtained from the univariate conventional MLE.

(2) The parameters of copula θ_C are obtained by maximizing the following expression:

$$\sum_{i=1}^{n} \log[c_Y(\widehat{F}_{Y_1}(y_{i,1}), \ldots, \widehat{F}_{Y_m}(y_{i,m})|\theta_C].$$

It should be noted that the above expression is obtained directly from the likelihood logarithm only by using marginal distributions in Step 1 and using the parameters of θ_C that were not estimated (Haugh (2016)). In fact, instead of estimating all the parameters, we used the empirical distribution function to estimate the community parameter.

4. Proposed Approach

In this method, we first identify and remove redundant variables using the copula function, and in the next step, we apply the LDA method to the remaining data and classify the data, because the LDA is a linearly controlled method. We expect it to perform better on residual data that have less structural correlation than the original data.

(1) The first step includes the following steps:
   (a) First, we fit Gaussian copula functions in pairs for the variables, then we estimate the copula function parameter using the pseudo-MLE method. By estimating the copula function parameter, we place the strongly correlated dimensions into a smaller subset. When the parameter ρ of the copula function for two continuous variables X_1 and X_2 is greater than 0.7, then the variables X_1 and X_2 are structurally strongly correlated, so they are included in the subset.
   (b) In the second step, we examine the dimensions in this subset and remove the dimensions that are linear combinations from other dimensions of this subset.
   (c) At the end, we select the variable that has the most variance in this subset and delete the rest of the variables. Because the behavior of the other variables is similar to the selected variable and we delete them as redundant variables.

(2) In the second step, we apply the LDA method to the remaining variables and classify the data.

5. Numerical results

In this section, we use a series of data including the classification of Covid-19 coughs using Mel-frequency coefficients for use in smart home appliances, which includes 26 different variables $X_1, ..., X_{26}$ and one classifier variable into two groups with Covid and not_Covid, which is available at `https://www.kaggle.com/datasets`. We divide this data into two sets of training and testing in the ratio of 70 to 30 (Dobbin and Simon (2011)).

In general, the performance of a classification method can be measured by different criteria. We use three common criteria: Precision, Recall, and Accuracy, which are defined based on True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) values ((Nutter et al. (2006); Metz (1978))):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}},$$

$$\text{P} = \frac{\text{TP}}{\text{TP+FP}},$$

$$\text{R} = \frac{\text{TP}}{\text{TP+FN}}.$$

Based on the proposed method, we first identify and remove redundant variables. To do this, first after estimating the Copula parameters and examining the variables that are structurally strongly

correlated with each other, the desired subsets are as follows:

$$\{X_3, X_4, X_5, X_6, X_8\},$$

$$\{X_2, X_7\}.$$

Now we check that these variables are not linear combinations of each other, i.e., for each i member we must have the above subsets: $\sum_i \alpha_i X_i = 0$, we have then $\alpha_i = 0$.

The study showed that all dimensions of this subset are linearly independent.

Finally, we select the variable that has more variance from these subsets and delete the other variables of these two subsets, so the reduced data is as follows:

$$(X_1, X_3, X_7, X_9, \ldots, X_{29}).$$

We now examine the LDA method for the main data and the reduced data and compare it with the common classification methods.

K-Nearest Neighbor (KNN) (Derrac et al. (2016)) proposes to sort each instance of the test based on its similarity to instances in the learning set and returns the most common class among these k instances (neighbors).

Naïve Bayesian (Saoudi et al. (2016)) applies Bayes' rule to detect the probability of a certain sample from a test sample depending to a particular class. The learning of this classifier is accomplished by computing the mean and variance of each dimension in each class.

Using the LDA method for Covid-19 data, the LDA coefficients diagram are as follows in Figure 1.

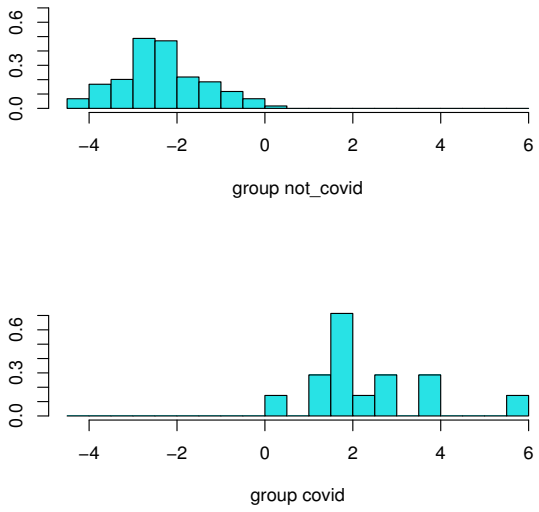

Figure 1. LDA coefficient for main data

And for reduced data we have as follows in Figure 2.

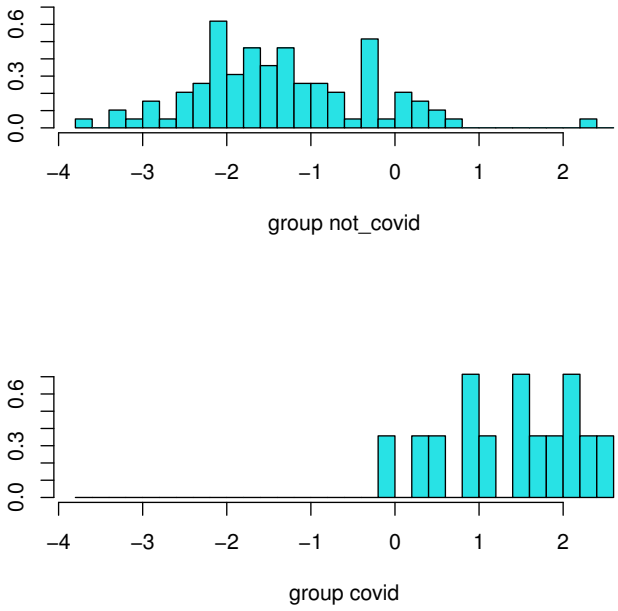

Figure 2. LDA coefficients for reduced data

Figures 1 and 2 show that by removing redundant variables, the coefficients LDA are more uniformly distributed between the variables, and Table 1 shows the values of different criteria for different classification methods.

Table 1. Values of different criteria for classification methods

Classification Method	Full Data			Reduction Data		
	Accuracy	P	R	Accuracy	P	R
LDA	0.89	0.96	0.9	1	1	1
Naïve Beyes	0.86	0.9	0.93	0.94	0.96	0.96
KNN	0.97	0.96	1	0.91	0.93	0.96

6. Conclusion

According to Table 1, we observed that after identifying and eliminating redundant variables using the copula function, the classification accuracy in the LDA method increased and the efficiency of this method for data classification improved.

As can be seen for the original data, the KNN method performs the classification better than the other two methods, but after identifying and eliminating the variables that are structurally corre-

lated, the LDA method, which is a linearly controlled method, works better and the accuracy of the classification increases. However, after removing redundant variables, the Naive Bayesian method also improves.

From the above, it follows that the LDA method is a more effective and efficient method for classifying data that are less correlated with each other or after identifying and eliminating variables that are highly structurally correlated (redundant data).

As a result, with these things said, the variable reduction method presented in this research is very suitable for big data analysis, and this method can be used to prepare data for different data analysis methods.

REFERENCES

Badakhshan Farahabadi, F., Fathi Vajargah, K. and Farnoosh, R. (2021). Dimension reduction big data using recognition of data features based on Copula function and principal component analysis, Adv. Math. Phys., Art. ID 9967368, pp. 8.

Derrac, J., Chiclana, F., García, S. and Herrera, F. (2016). Evolutionary fuzzy k-nearest neighbors algorithm using interval-valued fuzzy sets, Information Sciences, Vol. 329, pp. 144–163.

Dobbin, K.K. and Simon, R.M. (2011). Optimally splitting cases for training and testing high dimensional classifiers, BMC Medical Genomics, Vol. 4, No. 1, pp. 1–8.

Durante, F., Fernandez-Sanchez, J., and Sempi, C. (2013). A topological proof of Sklar's theorem, Applied Mathematics Letters, Vol. 26, No. 9, pp. 945–948.

Farag, A.A. and Elhabian, S. (2008). A tutorial on data reduction linear discriminant analysis (LDA), University of Louisville, Tech. Rep.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, Annals of Eugenics, Vol. 7, No. 2, pp. 179–188.

Haugh, M. (2016). An introduction to copulas, *IEOR E4602: Quantitative Risk Management*, lecture notes, Columbia University.

Lei, T., Lin, X.-H., and Sun, D.-W. (2019). Rapid classification of commercial cheddar cheeses from different brands using plsda, lda and spa–lda models built by hyperspectral data, Journal of Food Measurement and Characterization, Vol. 13, No. 4, pp. 3119–3129.

Li, Q. and Schonfeld, D. (2014). Multilinear discriminant analysis for higher-order tensor data classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36, No. 12, pp. 2524–2537.

Lopez-Paz, D., Hernández-Lobato, J. M., and Zoubin, G. (2013). Gaussian process vine copulas for multivariate dependence, in *International Conference on Machine Learning*, pp. 10–18, PMLR.

MacKenzie, D. and Spears, T. (2014). "The formula that killed Wall Street": The Gaussian copula and modelling practices in investment banking, Social Studies of Science, Vol. 44, No. 3, pp. 393–417.

Metz, C.E. (1978). Basic principles of roc analysis, Seminars in Nuclear Medicine, Vol. 8, pp. 283–298.

Nelsen, R.B. (2007). *An Introduction to Copulas*, Springer Science & Business Media.

Nutter, F.W., Esker, P.D. and Netto, R.A.C. (2006). Disease assessment concepts and the advancements made in improving the accuracy and precision of plant disease data, European Journal of Plant Pathology, Vol. 115, No. 1, pp. 95–103.

Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 10, No. 2, pp. 159–203.

Saoudi, M., Bounceur, A., Euler, R. and Kechadi, T. (2016). Data mining techniques applied to wireless sensor networks for early forest fire detection, in *Proceedings of the International Conference on Internet of Things and Cloud Computing*, pp. 1–7.

Zheng, D., Hong, Z., Wang, N. and Chen, P. (2020). An improved lda-based elm classification for intrusion detection algorithm in iot application, Sensors, Vol. 20, No. 6, pp. 1706.