Research Journal of Applied Sciences 11 (11): 1422-1426, 2016
ISSN: 1815-932X

Performance Analysis: An Integration of Principal Component Analysis and Linear Discriminant Analysis for a Very Large Number of Measured Variables

Hashibah Hamid, Fatinah Zainon and Tan Pei Yong
School of Quantitative Sciences, UUM College of Arts and Sciences,
University Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

Abstract: Principal Components Analysis (PCA) is a variable reduction technique helps to reduce a complex dataset to a lower dimensional subspace. This study is interested to investigate an approach for handling a problem occurred from considering a very large number of measured variables followed by a classification task. For such purpose, PCA has been used to extract and reduce of a very large number of variables that considered in the study. Then, a Linear Discriminant Analysis (LDA) which is commonly used for classification is constructed based on the reduced set of variables. The performance analysis of the constructed PCA+LDA was conducted and compared with the classical LDA Model using different size of sample (n) and different number of independent variables (p). The performance of PCA+LDA and classical LDA Model has been evaluated based on misclassification rate. The results demonstrated that PCA+LDA performed better than the classical LDA Model for small sample case. For large sample size case, PCA+LDA also performed better than the classical LDA especially when the measured independent variables is too large. The overall findings showed that the constructed PCA+LDA can be considered as a good approach for handling a very large number of measured variables and performing classification treatment.

Key words: LDA, PCA, classification, misclassification rate, large variables

INTRODUCTION

In general, classification methods not only play a role as classifiers but also are quantitative approaches for modeling qualitative responses. Thus, classification is a method to find mathematical relationships between a qualitative variable and a set of descriptive variables (Ballabio and Todeschini, 2009). However, the main objective of classification is to learn the discrimination model to achieve a minimum misclassification rate (Lee, 2010). It also helps in comparison between different sets of data and shows the difference points between agreement and disagreement. Classification allows analyzing the relationship between some characteristics and makes further statistical treatment.

According to Engle and Manganelli (2004) some existing classification methods can be grouped into three: parametric, non-parametric and semi-parametric methods. Parametric methods are stronger than nonparametric methods as it needs less data to make a strong conclusion (Neideen and Brasel, 2007). Nevertheless, all data points must exhibit a bell shaped curve which is normally distributed (Sheskin, 2004). Examples of parametric methods that frequently used are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Partial Least Squares Discriminant Analysis (PLSDA) and Soft Independent Modeling of Class Analogy (SIMCA).

On the other hand, a non-parametric method is known as distribution free (Higgins, 2004). Non-parametric methods allow more flexible methods than parametric in accommodating different distributions (Yu, 2012). Furthermore, it allows examining and introducing data without prior assumption about the data distribution (Kim, 2003). For example, Classification And Regression Trees (CART) and k-Nearest Neighbor (kNN) make no assumption about the distribution of the data.

Meanwhile, semi-parametric method is a combination of a parametric method and a non-parametric method. Semi-parametric method estimates the problem regardless of non-smooth measurement functions which consists of both infinite and finite dimensional unknown parameters and has very weak assumptions (Chen *et al*., 2003). Semi-parametric has advantage as it does not need any prior knowledge and information to model relation (Bauer, 2005). Logistic discriminant analysis is an example of semi-parametric method (Anderson, 1972).

Corresponding Author: Hashibah Hamid, School of Quantitative Sciences, UUM College of Arts and Sciences, University Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

However, this study focuses only on parametric method which is LDA due to it is as the most popular and important method in discriminant analysis and it has been widely used in many applications (Lu and Zhang, 2012). LDA works efficiently when the assumption of equal population covariance structures for groups are satisfied and the independent variables follow multivariate normal distribution (Okwonu and Othman, 2012).

Research problems: This study investigates an approach for tackling a problem of a very large number of measured variables and then executes classification task. Large variables will be highly computational and will suffer from the problem of curse of dimensionality. Curse of dimensionality occurs due to the rapid increase in the volume associated lead to the adding of extra dimensions to a mathematical space. The existence of many irrelevant variables leads to the problems of misclassification (Ball and Bruner, 2010). Furthermore, multicollinearity is always exist when the measured variables are large as they are highly correlated to each other (Mason and Brown, 1975). A high degree of multicollinearity will induce unacceptable in coefficient estimates (Voss, 2004). Thus, multicollinearity can cause some problems such as high standard errors and unreliable coefficients (Allison, 2012).

To solve these problems, variable reduction is needed to be performed (Samaneh *et al*., 2016). The most popular technique of variable reduction is Principal Components Analysis (PCA). PCA has been widely used in applications ranging from pattern recognition and multivariate time series prediction to visualization (Labib and Vemuri, 2006).

The main purpose of PCA is to investigate the underlying information from multivariate raw data. PCA helps to identify how to reduce a complex dataset to a lower dimension to show the hidden structures and simplify understanding (Shlens, 2014). By reducing the number of dimension, PCA helps in compress the data without much loss of information (Dash and Nayak, 2013).

After conducting PCA, LDA Model is constructed for classification purposes. LDA is commonly used for classification under parametric method (Hastie *et al*., 2009). LDA is an important classical statistical tool for verification, recognition and has been widely used in many applications (Kyperountas *et al*., 2005). LDA aims to classify objects into one of two or more predefined groups where the dependent variables are appearing in a qualitative form (Okwonu and Othman, 2012). It has been successfully used as a statistical tool of origin method in several classification problems (Qiao *et al*., 2008).

Therefore, this study intends to explore PCA to deal with a very large number of variables. Then, LDA Model is constructed to conduct classification task. The performance of the constructed PCA+LDA is examined in some data conditions. In a review on the past studies, it showed that this strategy had been conducted in many times. This implies that PCA and LDA are among the famous methods for variable reduction and classification. In fact, both of the PCA and LDA are still the interest of many researchers until now. Thus, this study is interested to integrate PCA and LDA for a very large number of measured variables and to perform classification task using the reduced set of variables resulting from PCA.

The main aim of this study is to perform variable reduction using PCA and to construct classification model based on LDA for classification purposes on a very large number of measured variables. To achieve this main goal, there are three specific objectives need to be fulfilled as:

- Performing PCA to extract and reduce of a very large number of measured variables
- Constructing LDA using the reduced set of variables obtained in step 1
- Comparing and evaluating the performances of the constructed PCA+LDA with classical LDA Model based on misclassification rates

MATERIALS AND METHODS

Data generation: Data are generated using R statistical software program. First, perform variable reduction using PCA, then construct LDA Model from the results of PCA. Next, compare and analyze the performance of the constructed PCA+LDA with classical LDA. Performance analysis is conducted using different sizes of sample (n) and different number of independent variables (p). According to Delaigle and Hall (2012), n = 30 and n = 100 can be considered as small and large sample sizes. The number of independent variables that are considered under n = 30 are p = 10, p = 20 and p = 30 while under n = 100, it was set to p = 10, p = 20, p = 30, p = 50, p = 70 and p = 90. These two different sample sizes with different number of independent variables are generated for the purpose of comparison in order to investigate and analyze the performance of the constructed PCA+LDA Model. All together we have nine simulation datasets for such intentions.

Procedures: This study involves few steps in order to integrate PCA and LDA as follows:

- Perform PCA to reduce a very large number of measured variables
- Estimate parameters; mean, covariance matrix and probability using the reduce set of variables

- Construct a classification model based on LDA using the estimated parameters
- Evaluate the performance of the constructed PCA+LDA Model based on minimum misclassification rate using

$$\sum_{k=1}^{n}\frac{error_k}{n}$$

where, k = 1, 2, …, n.

Computational of PCA: A data matrix is given with n sample size consisting of X_1, X_2, ..., X_p variables. The data is centered on the origin of the principal components (Y_p) and will influence neither the relationships of the data nor those variances of the variables. The first principal components, denoted as Y_1 is a linear combination of the variables X_1, X_2, ..., X_p in such a way that:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \quad (1)$$

and it can also be written in a matrix form as follows:

$$Y = \alpha^T X \quad (2)$$

The highest variance in the dataset will be accumulated in the first computed principal component (Y_1). Variance of Y_1 will become very large, due to large values of weights which are a_{11}, a_{12}, a_{1p}. To avoid this, weights are constrained by:

$$a_{11}^2 + a_{12}^2 + a_{1p}^2 = 1 \quad (3)$$

Similarly the second principal component (Y_2) is calculated. However, there is a condition that it is uncorrelated with Y_1 and the next highest variance accounted for. Y_2 can be written as:

$$Y_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \quad (4)$$

The calculation will be continued until p principal components are obtained which is equal to the original number of variables that are measured in a study. Moreover, the sum of variances of all variables will equal to the sum of variances of all principal components as the principle components explain all of the original information. Basically, PCA transforms all of the original variables to the principal components.

Only the important principal components will be chosen based on eigenvalues >1.0, following the work of Abdi and Williams (2010). As a result, a new set of variables which are reduced in size and uncorrelated to each other. Thus, these new set of variables are ready to be used for the construction of the PCA+LDA Model.

Construction of PCA+LDA Model: Suppose that $n_1$is observed from group 1 (π_1) and n_2 is observed from group 2 (π_2). For classification purposes based on LDA, the object is classified to π_1 if otherwise it will be classified to π^2:

$$(\mu_1 - \mu_2)^T \Sigma^{-1}\left[y_{PCA} - \frac{1}{2}(\mu_1 + \mu_2)\right] > \log\left(\frac{\rho_2}{\rho_1}\right) \quad (5)$$

Where:

μ_i = Mean vectors of π_1
y_{PCA} = Data vector to be classified after PCA procedure
Σ = Homogeneous covariance matrix
ρ_i = Prior probabilities of π_1 and i = 1, 2

The LDA Model is constructed from a reduced set of variables resulting from the procedures of PCA using following steps:

- Estimate means for π_1 and π_2 using $\mu_i = \Sigma y/n$ based on the reduced set of variables
- Compute homogeneous covariance matrix using

$$\Sigma = \frac{\Sigma(y_1 - \mu_1)(y_m - \mu_2)}{n-1}$$

- Estimate prior probabilities for π_1 and π_2 by $\rho_i = n_i/n$
- Construct LDA Model using Eq. 5
- Evaluate the performance of the constructed PCA+LDA Model based on the lowest misclassification rate through

$$\sum_{k=1}^{n}\frac{error_k}{n}$$

where, k = 1, 2, ..., n. The evaluation step is performed using the Leave-One-Out (LOO) procedure which is a well-known approximately unbiased method (Cawley, 2006). LOO also has the capability to gain maximum information from the datasets (Amendolia *et al*., 2003).

RESULTS AND DISCUSSION

The investigations based on different sample sizes (n) and different number of independent variables (p) are done to make comparison on the performance analysis of the constructed PCA+LDA with classical LDA Model.

Table 1: Comparison and performance analysis of the constructed PCA+LDA and classical LDA

Sample size: number of measured variables	Misclassification rate	
	PCA+LDA (misclassified object)	Classical LDA (misclassified object)
N = **30**		
p = 10	0.0333 (1)	0.0667 (2)
p = 20	0.0 (0)	0.2667 (8)
p = 30	0.0667 (2)	0.70 (21)
N = **100**		
p = 10	0.02 (2)	0.11 (11)
p = 20	0.0 (0)	0.02 (2)
p = 30	0.0 (0)	0.02 (2)
p = 50	0.0 (0)	0.0 (0)
p = 70	0.0 (0)	0.01 (1)
p = 90	0.01 (1)	0.22 (22)

Table 1 shows that under small sample size (n = 30), the performance of PCA+LDA is greatly improved than the classical LDA for all cases. The results indicate small misclassification rate under PCA+LDA compared to the classical LDA. In particular, for the case of n = 30 and p = 30, PCA+LDA obtains 6.67% misclassification rate while classical LDA achieves much higher misclassification rate with 70.0%. In other words, PCA+LDA has misclassified only 2 out of 30 objects while classical LDA misclassified 21 objects for the same data condition. For p = 20, PCA+LDA shows perfect classification but classical LDA gives 26.67% misclassification rate. These results demonstrated that PCA plays an important role in dealing with a large number of variables. If the size of sample is comparable to the number of variables, then it is useful to conduct PCA prior to make further analysis. It can be observed when p is closer to n, the misclassification rate will become higher. In general, the result has verified that PCA helps improved the misclassification rate.

For a large sample size (n = 100), there is almost no different in the performance between PCA+LDA and classical LDA Model. However, there are two cases where PCA+LDA is much better than the classical that is when the measured variables are very small (p = 10) or very large (p = 90). When p = 10, PCA+LDA gives only 2.0% misclassification rate while classical LDA provides higher misclassification rate with 11.0%. Meanwhile, when p = 90, PCA+LDA gives only 1.0% misclassification rate but classical LDA achieves much greater misclassification rate which is 22.0%.

When comparing the performance in terms of sample size, PCA+LDA and classical LDA performed better under larger sample size for most cases compared to small sample size. The discussion of the results can be summarized as follows:

- The relationship between difference sizes of p under small n (n = 30)
- When p gets larger, the misclassification rate also gets large for both PCA+LDA and classical LDA Model. However, PCA plays an important role as the performance of the constructed PCA+LDA is much better than the classical LDA Model, especially for large p
- The relationship between difference sizes of p under large n (n = 100)
- The performance of both PCA+LDA and classical LDA is quite consistent except for p = 10 and p = 90. In general, however, PCA+LDA performed better than the classical LDA in most cases especially for p = 90. This outcome shows that for a very large number of variables, PCA is good to be implemented before used in further analysis as it greatly helps to improve the model performance
- The relationship between n and p
- Regardless of whether the sample size is small or large, the performance of PCA+LDA is always better than the classical LDA especially when p is closer to n. PCA has been proven to reduce the misclassification rate of the constructed model
- The relationship between sample sizes
- Both PCA+LDA and classical LDA Models show better performance under large n compared to small n in most of the investigated cases. This outcome is consistent with the result of Dobbin *et al.* (2008), a large sample size will give more accurate results

CONCLUSION

In general, the entire results revealed that PCA+LDA performed better than the classical LDA Model for both cases of small and large sample sizes. The results also demonstrated that the nearer the number of independent variables to the sample size, the higher the misclassification rate for PCA+LDA and classical LDA Models. However, under n =1 00, the results are comparable between PCA+LDA and classical LDA but certain cases showed that PCA+LDA performed much better than the classical LDA model especially when p = 10 and p = 90. As a conclusion, PCA can be considered as a good technique to be used in dealing with very large number of variables before performing classification task especially for the case of large p and small n as well as for the case of p close to n.

ACKNOWLEDGEMENTS

The research would like to thank Ministry of Higher Education Research Grants and Universiti Utara Malaysia for financial support under Fundamental Research Grant Scheme (FRGS).

REFERENCES

Abdi, H. and L.J. Williams, 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat., 2: 433-459.

Allison, P., 2012. When can you safely ignore multicollinearity. Stat. Horiz., Vol. 5,

Amendolia, S.R., G. Cossu, M.L. Ganadu, B. Golosio and G.L. Masala *et al*., 2003. A comparative study of K-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening. Chemom. Intell. Lab. Syst., 69: 13-20.

Anderson, J.A., 1972. Separate sample logistic discrimination. Biometrics, 59: 19-35.

Ball, N.M. and R.J. Brunner, 2010. Data mining and machine learning in astronomy. Intl. J. Mod. Phys. D., 19: 1049-1106.

Ballabio, D. and R. Todeschini, 2009. Multivariate classification for qualitative analysis. Infrared Spectrosc. Food Qual. Anal. Control, 1: 83-104.

Bauer, D.J., 2005. A semiparametric approach to modeling nonlinear relations among latent variables. Struct. Equ. Model., 12: 513-535.

Cawley, G.C., 2006. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. Proceedings of the 2006 IEEE International Joint Conference on Neural Network Proceedings, July 16-21, 2006, IEEE, Norwich, England, ISBN:0-7803-9490-9, pp: 1661-1668.

Chen, X., O. Linton and V.I. Keilegom, 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica, 71: 1591-1608.

Dash, P. and M. Nayak, 2013. A study on principal component analysis for lossless data compression. Indian J. Res., 2: 125-129.

Delaigle, A. and P. Hall, 2012. Achieving near perfect classification for functional data. J. R. Stat. Soc. Ser. B., 74: 267-286.

Dobbin, K.K., Y. Zhao and R.M. Simon, 2008. How large a training set is needed to develop a classifier for microarray data?. Clin. Cancer Res., 14: 108-114.

Engle, R.F. and S. Manganelli, 2004. Quantile Prediction. In: Economic Forecasting, Elliott, G. and A. Timmermann (Eds.). Elsevier, Netherlands, pp: 964-968.

Hastie, T., R. Tibshirani and J. Friedman, 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edn., Springer, New York, pp: 520-528.

Higgins, J.J., 2004. An Introduction to Modern Nonparametric Statistics. Brooks/Cole, California, USA., ISBN:9780534387754, Pages: 366.

Kim, T.W., 2003. Nonparametric approaches for drought characterization and forecasting. BA Thesis, The University of Arizona, Tucson, Arizona. http://arizona.openrepository.com/arizona/handle/10150/280424.

Kyperountas, M., A. Tefas and I. Pitas, 2005. Methods for improving discriminant analysis for face authentication. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05), March 23, 2005, IEEE, Thessaloniki, Greece, ISBN:0-7803-8874-7, pp: 549-549.

Labib, K. and V.R. Vemuri, 2006. An application of principal component analysis to the detection and visualization of computer network attacks. Ann. Telecommun., 61: 218-234.

Lee, Y., 2010. Support Vector Machines for Classification: A Statistical Portrait. In: Statistical Methods in Molecular Biology, Heejung, B., X.Z. Kathy, L.V.E. Heather and M. Mazumdar (Eds.). Springer, Berlin, Germany, ISBN:978-1-60761-578-1, pp: 347-368.

Lu, Z. and Y. Zhang, 2012. An augmented lagrangian approach for sparse principal component analysis. Math. Program., 135: 149-193.

Mason, R. and W.G. Brown, 1975. Multicollinearity problems and ridge regression in sociological models. Soc. Sci. Res., 4: 135-149.

Neideen, T. and K. Brasel, 2007. Understanding statistical tests. J. Surg. Educ., 64: 93-96.

Okwonu, F.Z. and A.R. Othman, 2012. A model classification technique for linear discriminant analysis for two groups. Intl. J. Comput. Sci. Issues, 1: 125-128.

Qiao, Z., L. Zhou and J.Z. Huang, 2008. Effective linear discriminant analysis for high dimensional, low sample size data. Proceeding of the World Congress on Engineering, July 2-4, 2008, Texas A&M University, Cambridge, Massachusetts, ISBN:978-988-17012-3-7, pp: 2-4.

Samaneh, M.I., A. Khosro and Z. Leyla, 2016. Bayesian variable selection under collinearity of parameters. Res. J. Appl. Sci., 11: 428-438.

Sheskin, D., 2004. Handbook of Parametric and Nonparametric Statistical Procedures. 3rd Ed., Chapman and Hall /CRC, Boca Raton, ISBN: 9781584884408, Pages: 1193.

Shlens, J., 2014. A tutorial on principal component analysis. Comput. Sci., 1: 1-12.

Voss, D.S., 2004. Multicollinearity. Encycl. Soc. Meas., 2: 759-770.

Yu, Y., 2012. Bayesian and non-parametric approaches to missing data analysis. Ph.D Thesis, University of California, California, USA., https://escholarship.org/uc/item/3378b6tx.