JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Hamid, H., Mahat, N. I., & Ibrahim, S. (2021). Adaptive variable extractions with LDA for classification of mixed variables, and applications to medical data. *Journal of Information and Communication Technology, 20*(3), 305-327. https://doi.org/10.32890/jict2021.20.3.2

Adaptive Variable Extractions with LDA for Classification of Mixed Variables, and Applications to Medical Data

[1]Hashibah Hamid, [2]Nor Idayu Mahat & [3]Safwati Ibrahim
[1&2] School of Quantitative Sciences, Universiti Utara Malaysia, Malaysia
[3] Institute of Engineering Mathematics, Universiti Malaysia Perlis, Malaysia

hashibah, noridayu@uum.edu.my
safwati@unimap.edu.my

Received: 4/10/2020 Revised: 4/1/2021 Accepted: 7/1/2021 Published: 11/6/2021

ABSTRACT

The strategy surrounding the extraction of a number of mixed variables is examined in this paper in building a model for Linear Discriminant Analysis (LDA). Two methods for extracting crucial variables from a dataset with categorical and continuous variables were employed, namely multiple correspondence analysis (MCA) and principal component analysis (PCA). However, in this case, direct use of either MCA or PCA on mixed variables was impossible due to restrictions on the structure of data that each method could handle. Therefore, this paper executed some adjustments including a strategy for managing mixed variables so that those mixed variables were equivalent in value. With this, both MCA and PCA could be performed on mixed

variables simultaneously. The variables following this strategy of extraction were then utilised in the construction of the LDA model before applying them to classify objects going forward. The suggested models using three real sets of medical data were then tested, where the results indicated that using a combination of MCA and PCA for extraction and LDA could reduce the model's size. It had a positive effect on the model's classification task and better performance since it led towards minimising the leave-one-out error rate. Accordingly, the models proposed in this paper, including the strategy that was adapted, were successful in presenting good results over the full LDA model. Regarding the indicators that were used to extract and to retain the variables in the model, cumulative variance explained (CVE), eigenvalue, and a non-significant shift in the CVE (*constant change*) could be considered a useful reference or guideline for practitioners experiencing similar issues in future.

Keywords: Classification, linear discriminant analysis, multiple correspondence analysis, mixed variables, principal component analysis.

INTRODUCTION

Linear discriminant analysis (LDA) is frequently favoured in classification problems when explanatory variables have multivariate normal distribution, and the populations share an identical or uniform covariance matrix (Nazman & Erbas, 2017). The model tends to work in this case, even though the population deviates from normality (Gyamfi et al., 2017). However, notwithstanding this strength or robustness, LDA commonly experiences notable challenges, either when the objects (n) size is restricted if compared to the size of the variables (p), or when comparing to a similar number, n and p (Bodnar et al., 2020). As such, a singular issue in the model may be evident (Tharwat et al., 2017), may induce instability in the model itself, may produce a poor quality of the constructed model (Swesi & Bakar, 2019) or even worse, not possible to construct the model (An & Chen, 2009). As a result, accurate classification is doubtful.

On the other hand, the issue of managing such a condition could be overcome by altering the mathematical functions present in the LDA,

like easing or loosening its reliance on certain calculations when computing the inverse covariance matrix (Tarr et al., 2016). Despite this fact, limited research has addressed this possibility. A possible option in addressing this issue may include: (i) removing variables that are less informative in explaining the variations among the populations (*variable selection*); or (ii) mixing the initial variables by adopting a different approach in obtaining fewer new variables that have sufficient information (*variable extraction*) for classification purposes. Selecting variables is a reasonably straightforward method; nevertheless, it is sensitive of the correlation between the variables that may cause issues in the analysis, particularly with a vast number of variables (Zhang et al., 2017).

An alternative method, variable extraction, may address this issue. However, it requires a methodological approach in combining the initial variables to obtain the information considered important. Research examining the selection of variables and variable extraction in the area of classification has been carried out by numerous researchers, as reported in Peres and Fogliatto (2018), Ghosh and Shuvo (2019), as well as in AL-Jumaili (2020). However, the focus has been restricted to examining categorical or continuous variables only at one time. Therefore, it is important to explore the feasibility of variable extraction or varied types of data utilising LDA as the foundation in constructing the model. Currently, many problems tend to be associated with extracting enough data that are useful to merge or combine before constructing the LDA model.

Mixed-variable datasets are often unpredictable with various variable types, values, and structures. This challenge can be tackled either at data level or mathematical model level. The latter requires extra work but often preferred by most researchers. On the other hand, the former is much easier whereby pre-processing on data is often executed prior to complex analysis (Mohamed et al., 2018).

The objective of this paper is to explore two extraction methods as a pre-processing step prior to LDA, notably, multiple correspondence analysis (MCA), and principal component analysis (PCA) in order to extract important variables from the initial datasets, which are: (i) mixtures or combinations of continuous variables and categorical variables; and (ii) existence of correlation problems among the measured variables. The

proposed strategy begins by manipulating the measured mixed variable data in order to allow the use of PCA and MCA prior to the construction of LDA. PCA standardises all continuous variables and rates binary variables as -1 and 1, while MCA discretises all continuous variables based on their average values. Then, the extracted components from PCA and MCA that are free from correlation problems are used to construct LDA for classification purposes.

The difficulties that emerge are: (i) to ascertain whether the information or data is useful in order to minimise the proportion of errors in classifying objects via the LDA model; (ii) to place the process of variable extractions, model construction, and model evaluation in an appropriate way; and (iii) to ascertain if the combination of variable extractions and LDA is useful. The next part of this paper provides an outline of the notion surrounding LDA and variable extractions, which is then followed by outlining the suggested strategy in this approach. The results of the investigation are then presented, and lastly, the findings of this study are presented in the last section.

LINEAR DISCRIMINANT ANALYSIS WITH VARIABLE EXTRACTIONS

The main motivation of this paper is to adapt a strategy of variable extractions so that PCA and MCA can be implemented simultaneously when facing with a mixture of variables, in order to select only important variables to be included in the LDA model. The variables following the strategy of extractions are then used as input in building the classifier through LDA. Therefore, the variable extraction is a primordial step for automatic diagnosis, and the performance of the LDA model depends on the strategy used and the quality of the extracted variables.

Linear Discriminant Analysis

Let us first signify two groups as π_1 and π_2, whereby each group has: (i) a multivariate normal distribution with means, $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$, respectively; and (ii) a uniform covariance matrix, $\boldsymbol{\Sigma}$. The multivariate normal distribution of π_1 is $\mathbf{N}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$ and π_2 is $\mathbf{N}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$; thus the relative sizes of the posterior probabilities of the vector of measurements

($\boldsymbol{x}$) membership group, π_i, mentioned in Anderson (1958) can be expressed as in Equation 1:

$$\frac{f(2|\mathbf{x})}{f(1|\mathbf{x})} = \frac{\frac{\pi_2}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]}{\frac{\pi_1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]} \quad (1)$$

Next, given the assumption that the costs attributed to misclassification are the same, the upcoming object is assigned to π_1 if the relative is greater than 1, or else assigned to π_2. Some algebraic on Equation 2 gives:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] \ > \ 1 \quad (2)$$

where Equation 2 is recognised as an LDA function (Anderson, 1958).

However, from a practical perspective, the unknown parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are often substituted with the maximum likelihood estimators obtained from a randomly selected sample. A study by Alheety (2020) suggested that estimators are dependent on the sample relating to prior information. However, the estimation given by Equation 2 is deteriorated, especially when the variables are correlated, as shown by Krzanowski (1975; 1977). The larger the collinearity between the variables, the greater the loss in precision will be (Chandan et al., 1998). Serious stability issues will also eventuate if the data are highly multicollinear (Prats-Montalbán et al., 2006). Another issue is that the developed algorithm should first extract continuous variables or categorical variables when dealing with mixed variables. The applications of either PCA or MCA on mixed variables are inappropriate in this case to be performed simultaneously. Therefore, some adjustments are needed, including a strategy for managing mixed or combined variables using PCA and MCA simultaneously.

Principal Component Analysis with LDA

In this section, PCA is used to transform a group of initially related variables, $\boldsymbol{X} = (x_1, x_2, \ldots, x_m)$, into linear combinations of a group of

lessened pair-wise uncorrelated variables called principal components, $\boldsymbol{Z} = (z_1, z_2, \ldots, z_k)$, via linear combinations of $\boldsymbol{Z} = a^T \boldsymbol{X}$ for $k \leq m$ for (Jolliffe, 1986). Here, a^T is a matrix of eigenvalues, $\boldsymbol{a} = (a_{11}, a_{22}, \ldots, a_{kk})$, selected in giving a maximum variance of the elements of var$\boldsymbol{(Z)} = \boldsymbol{a^T \Sigma a}$ condition to $a_i^T a_i = 1$ and $a_i^T a_h$, where $i \neq h$.

PCA has been recognised for perceiving higher multivariate dimensional data into a smaller dimension, for example, two dimension (2D). This approach is useful if there is an indication of redundancy (correlation) amongst the variables. As such, this suggests that the independent variables are either near-linearly or linearly reliant on one another (Artoni et al., 2018), which may be due to displaying similar information. Moreover, utilising PCA in classification has been proven workable in facial recognition problems by Barnouti et al. (2016) and Deshpande and Ravishhankar (2017), and has consequently become a notable option in simplifying data before constructing the classification model (i.e. Jamal et al., 2018; Li, 2017; Nasution et al., 2018). The research undertaken by these scholars indicated that effort was focused on the application of PCA to reduce the range of variables prior to classification. Nevertheless, the majority of dialogue in these studies was restricted to continuous variables. Therefore, this research explores how PCA can be employed for mixed variables.

Multiple Correspondence Analysis with Classification

To enhance the PCA method, MCA was initially developed between the early 1960s and late 1970s when the former method was unable to estimate optimum properties and offer compelling tools in depicting the hidden structure in a set of categorical variables (Hamid et al., 2018). The initial creation in Guttman (1941) showed that the MCA method was noted as the PCA of qualitative or nominal variables. As highlighted by Josse and Husson (2016), MCA can be inferred as PCA using categorical data as found in survey-related studies, in which most of the information was categorical. Here, MCA projects q categorical data into a relatively smaller subspace (s), which accounts for the maximum variance of the categorical data (Kaminska et al., 1999). Similar to PCA, the initial dimension describes the greatest variance in the data, while the second dimension depicts a maximum of the remaining variation and so forth (Blasius & Thiessen, 2000). The capability of MCA in simplifying multiple relationships between

categorical variables also helps scholars and other experts who are keen on performing classification activities and need to handle a vast range of variables (Das & Sun, 2016; Das et al., 2018; Sivasankaran & Balasubramanian, 2020).

Nonetheless, those experts or practitioners employing MCA will often experience issues such as the intricacy and shortage of data when examining numerous categorical variables at the same time (Mori et al., 2016). Therefore, the results would fail to be well understood, given the perception of numerous variables would be difficult to comprehend. The entire process of analysis would also become quite arduous and complicated in the situation where the data were sparse (Messaoud et al., 2007). However, one approach in preventing issues of this kind from eventuating is by using subsets or combining some (Mori et al., 2016), thereby providing an option in the selection process that is both rational and easily explained.

Moreover, it could assist in simplifying the structures that the analysis is trying to explain. According to Ali et al. (2018), a relatively low dimensional view via MCA could represent the categorical variables. Likewise, as highlighted by D'Enza and Greenacre (2012), MCA seeks to detect a lesser set of artificial dimensions through maximising the explained variability of the categorical data. Indeed, prior research has revealed that MCA is capable of dealing with categorical variables in trying to determine patterns and correlations among the data (see Das et al., 2018; Dungey et al., 2018; Zhang et al., 2017). Therefore, it is useful for experts who are keen to reduce the dimension of the data when faced with numerous categorical variables. However, at this stage, scant interest has been shown by scholars when examining mixed variables. The interest of this current research is towards adapting MCA to reduce the size of measured q categorical variables into s extracted components in relation to the LDA model.

Employing PCA or MCA will help in addressing classification issues quite simply, as PCA or MCA can be used to decrease the number of initial variables to fewer extracted components. In this way, the extracted components can be utilised to build the classification model. However, while this process may seem appropriate, consideration needs to be given on the selection of components to support the objective in building the model. PCA and MCA both extract components that

depict the unpredictability of the initial variables. Nonetheless, LDA requires sound and good discriminators to confirm that it functions optimally for future classification needs.

METHODOLOGY

This paper presents a strategy by manipulating two extraction methods, PCA and MCA, on mixed continuous and categorical variables aimed at revealing meaningful structures in multivariate data. The adapted strategy aimed to increase feasibility of implementing both PCA and MCA simultaneously on the mixed variables. Therefore, the construction of classifiers using a single model, LDA, with different types of variables beforehand could be implemented ideally. The constructed models were then tested on three medical datasets that have various types of variables for validation purposes.

Model Construction and Evaluation

Classification of data with numerous and mixed variables was undertaken in this study by employing three main processes: (i) extracting the variables utilising PCA or MCA; (ii) building the LDA model by employing the extracted components; and (iii) evaluating the built model by computing the misclassification rate and explaining the percentage of objects that were misclassified. Having the extracted $\boldsymbol{Z} = (z_1, z_2, \ldots, z_k)$ components from PCA and/or MCA, then the classification rule in Equation 3 is adjusted by:

$$(\hat{\boldsymbol{z}}_{\boldsymbol{1}} - \hat{\boldsymbol{z}}_{\boldsymbol{2}})^T \boldsymbol{S}_{\boldsymbol{K}}^{-\boldsymbol{1}} \left[z - \frac{1}{2}(\hat{\boldsymbol{z}}_{\boldsymbol{1}} + \hat{\boldsymbol{z}}_{\boldsymbol{2}}) \right] > 1 \qquad (3)$$

where $\hat{z}_1$ and $\hat{z}_2$ are the estimated mean vectors of $\mathbf{Z}$ components in Group 1 and Group 2, respectively, and $\boldsymbol{S}_k^{-1}$ is the inverse of estimated uniform covariance matrix $\mathbf{Z}$. The classification processes were organised in a leave-one-out manner upon which n-1 objects were utilised to extract the variables and to build the LDA model. Then, the built model was tested on the excluded objects. The leave-one-out was selected to prevent any bias on the built LDA model. Algorithm 1 lists the details of the steps and processes that were carried out.

Algorithm 1: Model Construction and Evaluation

Step 1: Prepare the data $\boldsymbol{X}$ for variable extraction using

1.1 PCA: standardise all continuous variables using z-score and rated values of binary variables as either -1 or 1.

1.2 MCA: for each value of continuous variable, do

$$x_{ij} = \begin{cases} 0 & if \quad x_{ij} < \bar{x}_j \\ 1 & if \quad x_{ij} \geq \bar{x}_j \end{cases}$$

where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.

Step 2: Omit one object from a sample (n), $\boldsymbol{X}_i$.

Step 3: Perform PCA (and MCA) using the remaining $n - 1$ objects ($\boldsymbol{X}_{-i}$) to extract fewer $\mathbf{Z}$ components than the number of variables in $\boldsymbol{X}_{-i}$.

Step 4: Estimate all parameters $\hat{z}_1$, $\hat{z}_2$ and $\boldsymbol{S}_k^{-1}$ based on the extracted components $\mathbf{Z}$ produced in Step 3.

Step 5: Construct LDA models as in Equation 3 based on the estimated parameters in Step 4.

Step 6: Estimate the group of the omitted object $\boldsymbol{X}_i$ from Step 2.

Step 7: Compare the estimated group and the actual group of the omitted object. If there is a difference, count error as 1, otherwise count as 0.

Step 8: Repeat Steps 2–7 until all objects have been omitted in turn.

Step 9: Compute the error rate by the total number of error over the size of sample.

In summary, this paper recommends two LDA models: (i) one model that combines LDA with PCA (LDA+PCA); and (ii) one model that combines LDA and MCA (LDA+MCA). The LDA model involving all variables (full LDA model or original LDA model) is also depicted in determining if the recommended models are sound, acceptable, and can be adapted.

Variable Extraction with Mixed Variables

As depicted in Algorithm 1, first, the measured mixed variables were manipulated accordingly to allow for the feasibility use of PCA and MCA, by following some practices earlier performed by Krzanowski (1975; 1977) and Mahat et al. (2007). This strategy aimed to avoid greater misleading in implementing both PCA and MCA. In PCA, the standardisation of continuous variables using *z*-score control range values of input variables, while setting the binary values as either -1 or 1 was simply to create dissimilarity among the values. Meanwhile, for MCA, categorical variables remained as they were but discretisation was performed on continuous variables as a way to transformed continuum values to categorical.

Next, PCA or MCA was employed to extract relevant information from sets of manipulated data, containing both continuous and categorical variables. The selection of extracted components was important in this case, given that it could influence the performance of the LDA model, and at the same time, the LDA model would be free from correlated components. In testing this aspect, the current study examined three indicators: (i) eigenvalue, (ii) cumulative variance explained (CVE), and (iii) constant change in the CVE, to confirm their effects on minimising the error rate.

This paper extracted the components with an eigenvalue of at least 1.0 following Kaiser (1960). However, while the selection that was based on CVE was subjective, it was agreed to adopt it (Stevens, 2002) in order to extract the components with at least 70 percent and 80 percent of the variance explained. Moreover, while the eigenvalue might be quite inflexible, CVE could be fairly subjective. Therefore, this paper also set out to extract components as long as the incremental number of components offered no significant change regarding the CVE. Here, the difference in size needed to be smaller than unity.

Medical Datasets

The models that were adapted, as suggested above, were tested on three real sets of medical data that had various variable types such as *full* and *reduced sets of breast cancer* and *heart disease* taken from Krzanowski (1975, 1980) and Mahat et al. (2007). Besides, these data

were reported to have great correlation among the variables, hence dealing with such threat is a must. The breast cancer dataset represented 137 women diagnosed with breast cancer (tumours), 78 of the cases were benign (π_1) and 59 were malignant (π_2). Regarding the variables, the original data consisted of 15 variables, where six were ordinal variables each with a score ranging between 0 and 10, four nominal variables with three conditions each, three binary variables, and two continuous variables. Data manipulation was made to fit the proposed models where the ordinal variables were treated as a continuous form, while the nominal variables were converted to binary values giving a fresh set of data consisting of eleven binary variables and eight continuous variables. The lessened breast cancer data were attained by converting the initial nominal and ordinal variables into binary and continuous types; consequently, yielding six binary variables and seven continuous variables. The heart data represented 270 patients, consisting of 16 variables, of which 120 patients were diagnosed with heart disease (π_1) and the remainder were void of heart illness (π_2). The initial dataset consisted of three nominal variables (three states), three binary, and seven continuous variables. The nominal variables were treated as binary variables, resulting in a fresh dataset consisting of nine binary variables and seven continuous variables. Generalising this strategy to a categorical variable with multistate would not be a problem as one could create *q-1* dummy binary variables, where *q* represented the category state of the respective categorical variable.

Since the measured variables were mixed, the extracted components that contained high loading would be more from the continuous variables as compared to the binary variables, given that the applications of PCA on the mixed variables were not compatible to be performed simultaneously. This was also due to issues relating to domination and variability of the continuous variables being significantly higher as compared to the binary variables (Hamid et al., 2017). The same problem was faced by Vyas and Kumaranayake (2006) who derived the indices of socio-economic status involving mixed variables issues (i.e. binary variables derived from categorical variables), whereby PCA was employed to diminish the dimensionality of the data. Once again, the issue of variability and domination of the continuous variables surfaced, indicating that PCA was inappropriate to be used for mixed data types.

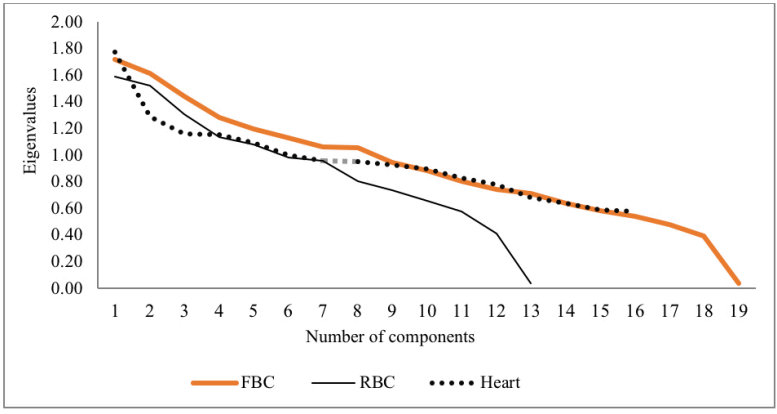
RESULTS AND ANALYSIS

Two classification models were proposed by merging (i) LDA and PCA and (ii) LDA and MCA on mixed variables simultaneously, with the new strategy mentioned. The extraction methods in selecting the components used three indicators: (i) eigenvalue, (ii) CVE, and (iii) insignificant change in the CVE (*constant change*). Following the proposed models, this study investigated the capability of the suggested strategy in managing a number of mixed variables for the purpose of overcoming classification problems.

Figure 1 illustrates the approach that was adopted to investigate the eigenvalues of PCA for the three medical datasets. As can be seen in the figure, the eigenvalues decreased as the number of components grew. The performance here showed the significance of PCA, in which the initial extracted component signified the maximum variance in the data, trailed by the second extracted component, and so forth. Nevertheless, it is hard to observe or distinguish an 'elbow' from the lines, which presented challenges in confirming the optimal number of extracted components. The benchmark used to extract the components with an eigenvalue of at least 1.0 produced eight extracted components for the full breast cancer data and five extracted components for both reduced breast cancer and heart datasets.

Figure 1

Eigenvalues Against the Number of Components in PCA



Figures 2 to 4 show the results of examining CVE in PCA and MCA for all medical datasets. The curve representing the CVE in MCA can be seen in Figure 2 as being marginally higher as compared to the CVE curve in PCA for the initial ten extracted components. It can then be seen that both curves began to become nearer when more extracted components came into play. Figure 3 depicts a similar behaviour of the CVE by PCA and MCA for reduced breast cancer data. The gap seen among both curves was evident for the initial seven extracted components, with the curves becoming nearer once more components were extracted from the data. Lastly, Figure 4 shows the CVE for the heart dataset. It can be seen that the CVE in MCA was marginally higher as compared to the CVE in PCA for a majority of components extracted.

Figure 2

Cumulative Variance Explained in PCA and MCA for Full Breast Cancer Data

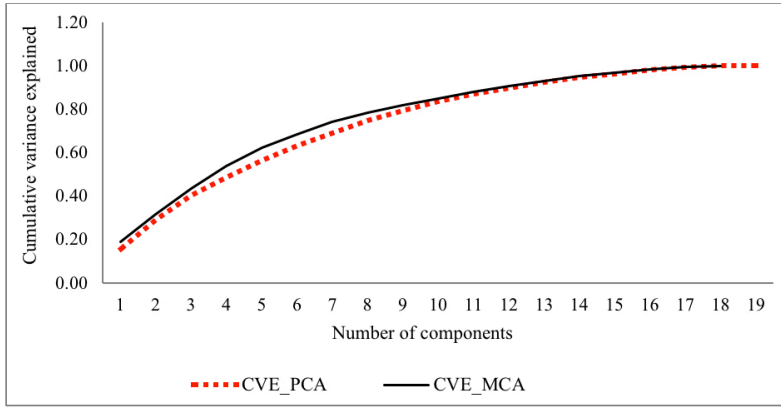


Figure 3

Cumulative Variance Explained in PCA and MCA for Reduced Breast Cancer Data

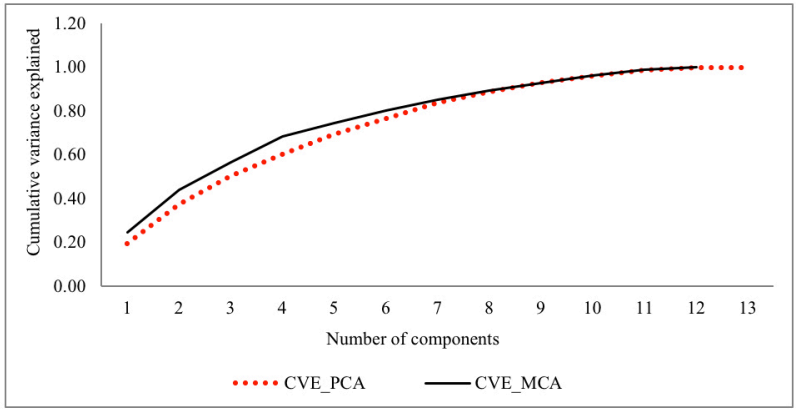
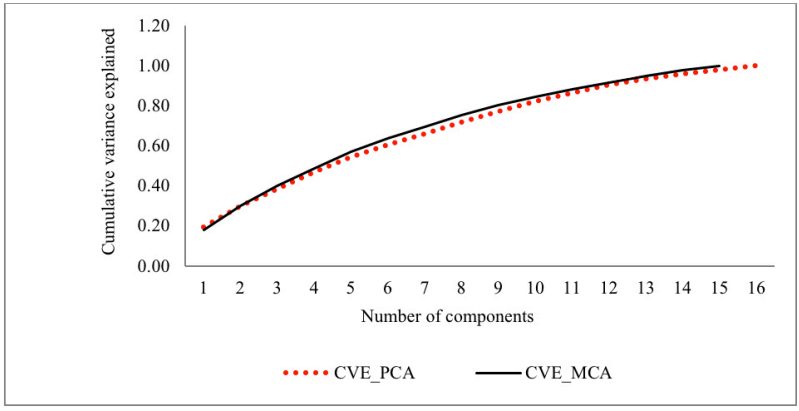


Figure 4

Cumulative Variance Explained in PCA and MCA for Heart Data



Reviewing the constant change in the CVE for all data as illustrated in Figures 5 to 7, this indicated the declining magnitude of change as the quantity of extracted components increased. This 'monotonic decrement' supported the CVE results, as shown in Figures 2 to 4. Figures 5 to 7 also demonstrate a minor difference between the results of PCA and MCA. It is shown that PCA and MCA were inclined to offer comparable results in extracting the components from the initial sets of data.

Figure 5

Constant Changes in CVE in PCA and MCA for Full Breast Cancer Data

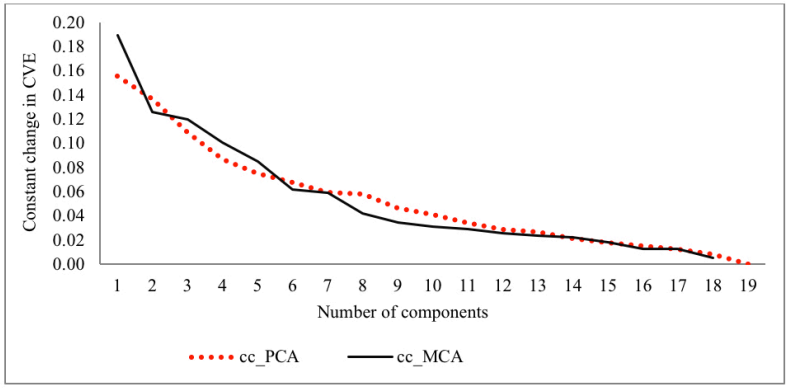


Figure 6

Constant Changes in CVE in PCA and MCA for Reduced Breast Cancer Data

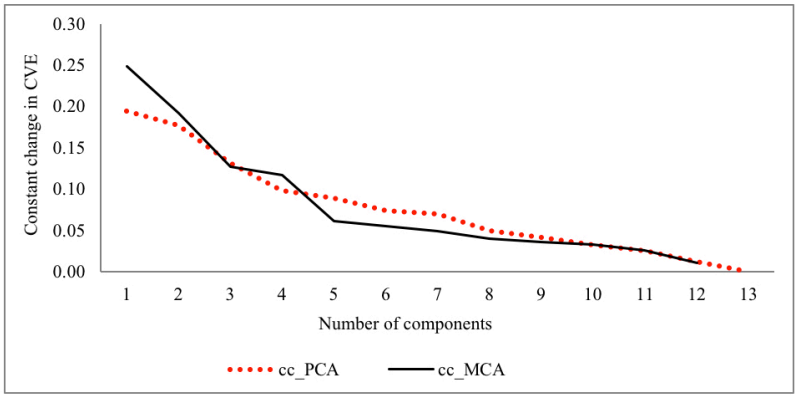
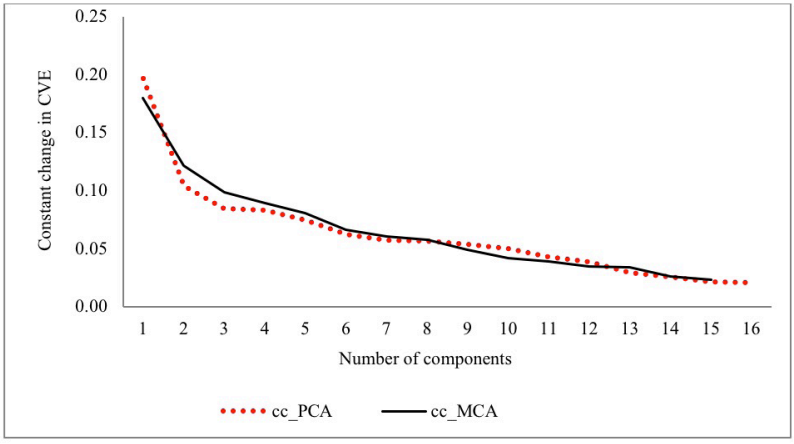


Figure 7

Constant Changes in CVE in PCA and MCA for Heart Data



A summary of the performance of the suggested models based on the leave-one-out error rate is presented in Table 1. The table depicts the best choice regarding the number of components based on the

eigenvalue, CVE at 70 percent and 80 percent, and the constant change in the CVE for the three sets of data, namely heart data, reduced breast cancer (RBC) data, and full breast cancer (FBC) data. The respective leave-one-out error rate was calculated to measure the performance of the proposed models and included the performance of the full LDA model for the purpose of comparison. For the FBC dataset, the LDA model with all 19 measured variables (original LDA model) resulted in a value close to 10 (error rate = 0.07) that represented misclassified patients. On the other hand, better performance was seen in the proposed LDA+PCA model for all indicators that were used. Meanwhile, the suggested LDA+MCA model exhibited a near-perfect classification. Here, both of the proposed models demonstrated an improvement over the full LDA model (original model). Furthermore, the models that used constant change of CVE extracted the smallest number of components, possibly indicating that this indicator was reasonably better than other indicators.

Likewise, the RBC dataset showed an error rate of 0.04 (six patients not classified correctly) depicted by using the full LDA model. In spite of that, the least chosen components were four and five, respectively, for the LDA+PCA and LDA+MCA models, which were far less than the full LDA model. These fewer variables successfully offered better performance on the classification result. The final dataset (heart data) showed that the full LDA model had an error rate of 0.08 (22 patients were classified incorrectly) when using all 16 variables in constructing the LDA model. Nevertheless, the LDA models with extracted components using either PCA or MCA proved to be far superior, producing a smaller error rate for all indicators used.

Generally, the results obtained on the leave-one-out error rate revealed that the proposed LDA+PCA and LDA+MCA models were significantly enhanced as compared to the full LDA model. The models showed good achievement with fewer variables (components) at hand.

Table 1

Results of LDA with Two Extraction Methods for All Three Real Medical Datasets

Classification models	Extraction indicators	Full breast cancer data		Reduced breast cancer data		Heart data	
		Number of selected components	Error rate	Number of selected components	Error rate	Number of selected components	Error rate
Full LDA (original model)	--	19	0.07	13	0.04	16	0.08
LDA + PCA	Eigenvalue	8	0.01	5	0.01	5	0.01
	CVE 70%	8	0.01	6	0.01	8	0.00
	CVE 80%	10	0.00	7	0.01	10	0.01
	Constant change	5	**0.00***	4	**0.01***	3	**0.00***
LDA + MCA	CVE 70%	7	0.00	5	**0.01***	8	0.01
	CVE 80%	9	0.00	6	0.01	9	0.01
	Constant change	6	**0.00***	5	**0.01***	3	**0.01***

* values with bold represent the best results

CONCLUSION

In this paper, the use and application of variable extraction methods (i.e. PCA and MCA) have been demonstrated in solving problems associated with classification tasks. The methods of extraction were shown to offer a major decrease in the number of variables used, which is useful in constructing the LDA model. Furthermore, the findings of LDA using the two extraction methods (i.e. proposed models) for all datasets were better as compared to the original model (full LDA). Accordingly, this infers that PCA and MCA could both be

utilised as alternate solutions in reducing the number of continuous and categorical variables when addressing classification challenges, even when facing mixed variables.

Similarly, the indicators used in this paper provided a differing number of extracted components, although their performance in classifying was similar. The findings based on the results confirmed that using those three indicators could affect the performance of LDA and as such, was considered reliable. However, no indicator can be viewed as the best and therefore, the use of such indicators should be used cautiously in line with the objectives and requirements of the investigation as well as with the structure of the model. According to Krzanowski (1987), it is acceptable to subjectively select variables from the initial data used as long as the degree of variation maintained by the reduced components is satisfactory. As a whole, both models, as proposed in this paper, offer suitable options to be applied in practice for the classification purposes primarily once facing variables that are correlated. Moreover, the strategy adopted in handling mixed variables is appropriate in offering good classification outcomes.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Alheety, M. (2020). New versions of liu-type estimator in weighted and non-weighted mixed regression model. *Baghdad Science Journal, 17*(1(Suppl.), 0361. http://bsj.uobaghdad.edu.iq/index.php/BSJ/article/view/ 5022

Ali, F., Dissanayake, D., Bell, M., & Farrow, M. (2018). Investigating car users' attitudes to climate change using multiple correspondence analysis. *Journal of Transport Geography, 72*, 237-247. https://doi.org/10.1016/j.jtrangeo.2018.09.007

AL-Jumaili, A. A. (2020). Hybrid method of linguistic and statistical features for Arabic sentiment analysis. *Baghdad Science Journal, 17*(1(Suppl.), 0385. https://doi.org/10.21123/bsj.2020.17.1(Suppl.).0385

An, J., & Chen, Y. P. P. (2009). Finding rule groups to classify high dimensional gene expression datasets. *Computational Biology and Chemistry, 33*, 108-113. https://doi.org/10.1016/j.compbiolchem.2008.07.031

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: John Wiley & Sons, Inc.

Artoni, F., Delorme, A., & Makeig, S. (2018). Applying dimension reduction to EEG data by principal component analysis reduces the quality of its subsequent independent component decomposition. *Neurolmage, 175*, 176-187. https://doi.org/10.1016/j.neuroimage.2018.03.016

Barnouti, N. H., Al-Dabbagh, S. S., Matti, W. E., & Naser, M. A. (2016). Face detection and recognition using Viola-Jones with PCA-LDA and square Euclidean distance. *International Journal of Advanced Computer Science and Applications (IJACSA), 7*(5), 371-377. https://doi.org/10.14569/ijacsa.2016.070550

Blasius, J., & Thiessen, V. (2000). Methodological artifacts in measures of political efficacy and trust: A multiple correspondence analysis. *Political Analysis, 9*(1), 1-20. https://doi.org/10.1093/oxfordjournals.pan.a004862

Bodnar, T., Mazur, S., Ngailo, E., & Parolya, N. (2020). Discriminant analysis in small and large dimensions. *Theory of Probability and Mathematical Statistics*, (100), 21-41. https://doi.org/10.1090/tpms/1096

Chandan, M., White, H., & Wuyts, M. (1998). *Econometrics and data analysis for developing countries*. London: Routledge.

Das, S., & Sun, X. (2016). Association knowledge for fatal run-off-road crashes by multiple correspondence analysis. *IATSS Research, 39*(2), 146-155. https://doi.org/10.1016/j.iatssr.2015.07.001

Das, S., Avelar, R., Dixon, K., & Sun, X. (2018). Investigation on the wrong way driving crash patterns using multiple correspondence analysis. *Accident Analysis & Prevention, 111*, 43-55. https://doi.org/10.1016/j.aap.2017.11.016

D'Enza, A. I., & Greenacre, M. J. (2012). Multiple correspondence analysis for the quantification and visualization of large categorical data sets. In A. Di Ciaccio, M. Coli & J. M. A. Ibaňez (Eds.), *Advanced statistical methods for the analysis of large data-sets: Studies in theoretical and applied statistics* (pp. 453-463). Berlin Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-21037-2_41

Deshpande, N. T., & Ravishhankar, S. (2017). Face detection and recognition using Viola-Jones algorithm and fusion of PCA and ANN. *Advances in Computational Sciences and Technology, 10*(5), 1173-1189.

Dungey, M., Tchatoka, F. D., & Yanotti, M. B. (2018). Using multiple correspondence analysis for finance: A tool for assessing financial inclusion. *International Review of Financial Analysis, 59*, 212-222. https://doi.org/10.1016/j.irfa.2018.08.007

Ghosh, J., & Shuvo, S. B. (2019). Improving classification model's performance using linear discriminant analysis on linear data. In *the 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2019* July 6 (pp. 1-5). IEEE. https://doi.org/10.1109/icccnt45670.2019.8944632

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst, P. Wallin & L. Guttman (Eds.), *The prediction of personal adjustment* (pp. 319-348). New York: Social Science Research Council.

Gyamfi, K. S., Brusey, J., Hunt, A., & Gaura, E. (2017). Linear classifier design under heteroscedasticity in linear discriminant analysis. *Expert Systems with Applications, 79*, 44-52. https://doi.org/10.1016/j.eswa.2017.02.039

Hamid, H., Mei, L. M., & Yahaya, S. S. S. (2017). New discrimination procedure of location model for handling large categorical variables. *Sains Malaysiana, 46*(6), 1001-1010. https://doi.org/10.17576/jsm-2017-4606-20

Hamid, H., Ngu, P. A., & Alipiah, F. M. (2018). New smoothed location models integrated with PCA and two types of MCA for handling large number of mixed continuous and binary variables. *Pertanika Journal of Science & Technology, 26*(1), 247-260.

Jamal, A., Handayani, A., Septiandiri, A. A., Ripmiatin, E., & Effendi, Y. (2018). Dimensionality reduction using PCA and K-means clustering for breast cancer prediction. *Lontar Komputer: Jurnal Ilmiah Teknologi Imformasi,* 192-201. https://doi.org/10.24843/lkjiti.2018.v09.i03.p08

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software, 70*(1), 1-31. https://doi.org/10.18637/jss.v070.i01

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151. https://doi.org/10.1177/001316446002000116

Kaminska, A., Ickowicz, A., Plouin, P., Bru, M. F., Dellatolas, G., & Dulac, O. (1999). Delineation of cryptogenic lennox–gastaut syndrome and myoclonic astatic epilepsy using multiple correspondence analysis. *Epilepsy Research, 36*, 15-29. https://doi.org/10.1016/s0920-1211(99)00021-2

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association, 70*(352), 782-790. https://doi.org/10.1080/01621459.1975.10480303

Krzanowski, W. J. (1977). The performance of fisher's linear discriminant function under non-optimal conditions. *Technometrics, 19*, 191-200. https://doi.org/10.1080/00401706.1977.10489527

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics, 36*, 493-499. https://doi.org/10.2307/2530217

Krzanowski, W. J. (1987). Selection of variables to preserve multivariate data structure using principal components. *Applied Statistics, 36*(1), 22-33. https://doi.org/10.2307/2347842

Li, M. (2017). Application of cart decision tree combined with PCA algorithm in intrusion detection. In *the 8th International Conference on Software Engineering and Service Science (ICSESS)* Nov 24 (pp. 38-41). IEEE. https://doi.org/10.1109/icsess.2017.8342859

Mahat, N. I., Krzanowski, W. J., & Hernandez, A. (2007). Variable selection in discriminant analysis based on the location model for mixed variables. *Advances in Data Analysis and Classification, 1*(2), 105-122. https://doi.org/10.1007/s11634-007-0009-9

Messaoud, R. B., Boussaid, O., & Rabaséda, S. L. (2007). A multiple correspondence analysis to organize data cubes**.** *Databases and Information Systems IV: Frontiers in Artificial Intelligence and Applications, 155*(1), 133-146.

Mohamed, R., Zainudin, M. N. S., Sulaiman, M. N., Perumal, T., & Mustapha, N. (2018). Multi-label classification for physical activity recognition from various accelerometer sensor positions. *Journal of Information and Communication Technology, 17*(2), 209–231. https://doi.org/10.32890/jict2018.17.2.3

Mori, Y., Kuroda, M., & Makino, N. (2016). Sparse multiple correspondence analysis: Nonlinear principal component analysis and its applications. Singapore: Springer. https://doi.org/10.1007/978-981-10-0159-8_5

Nasution, M. Z. F., Sitompul, O. S., & Ramli, M. (2018). PCA based feature reduction to improve the accuracy of decision tree C4.5 classification. *Journal of Physics: IOP Conference Series, 978*, 012058. https://doi.org/10.1088/1742-6596/978/1/012058

Nazman, E., & Erbas, S. (2017). Evaluation of group homogeneity in Gaussian mixture models using combined cluster and discriminant analysis. *Sinop University Journal of Natural Sciences, 2*(1), 121-132.

Peres, F. A., & Fogliatto, F. S. (2018). Variable selection methods in multivariate statistical process control: A systematic literature review. *Computers & Industrial Engineering, 115*, 603-619. https://doi.org/10.1016/j.cie.2017.12.006

Prats-Montalbán, J. M., Ferrer, A., Malo, J. L., & Gorbeña, J. A. (2006). Comparison of different discriminant analysis techniques in a steel industry welding process. *Chemometrics and Intelligent Laboratory Systems, 80*, 109-119. https://doi.org/10.1016/j.chemolab.2005.08.005

Sivasankaran, S. K., & Balasubramanian, V. (2020). Investigation of pedestrian crashes using multiple correspondence analysis in India. *International Journal of Injury Control and Safety Promotion, 27*(2), 144-155. https://doi.org/10.1080/17457300.2019.1681005

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences (4th edition)*. New Jersey: Lawrence Erlbaum Associates, Inc.

Swesi, I. M. A. O., & Bakar, A. A. (2019). Feature clustering for PSO-based feature construction on high-dimensional data. *Journal of Information and Communication Technology, 18*(4), 439-472. https://doi.org/10.32890/jict2019.18.4.3

Tarr, G., Müller, S., & Weber, N. C. (2016). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis, 93*, 404-420. https://doi.org/10.1016/j.csda.2015.02.005

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications, 30*(2), 169-190. https://doi.org/10.3233/aic-170729

Vyas, S., & Kumaranayake, L. (2006). Constructing socio-economic status indices: How to use principal components analysis. *Health Policy and Planning, 21*(6), 459-468. https://doi.org/10.1093/heapol/czl029

Zhang, D. F., Chen, Y. C., Chen, H., Zhang, W. D., Sun, J., Mao, C. N., Su, W., Wang, P., & Yin, X. (2017). A high-resolution MRI study of relationship between remodelling patterns and ischemic stroke in patients with atherosclerotic middle cerebral artery stenosis. *Frontiers in Aging Neuroscience, 9*, 140. https://doi.org/10.3389/fnagi.2017.00140