Sains Malaysiana 46(6)(2017): 1001–1010
http://dx.doi.org/10.17576/jsm-2017-4606-20

New Discrimination Procedure of Location Model for Handling Large Categorical Variables

(Prosedur Diskriminasi Baharu Model Lokasi untuk Mengendalikan Pemboleh Ubah Kategori Besar)

HASHIBAH HAMID*, LONG MEI MEI & SHARIPAH SOAAD SYED YAHAYA

ABSTRACT

The location model proposed in the past is a predictive discriminant rule that can classify new observations into one of two predefined groups based on mixtures of continuous and categorical variables. The ability of location model to discriminate new observation correctly is highly dependent on the number of multinomial cells created by the number of categorical variables. This study conducts a preliminary investigation to show the location model that uses maximum likelihood estimation has high misclassification rate up to 45% on average in dealing with more than six categorical variables for all 36 data tested. Such model indicated highly incorrect prediction as this model performed badly for large categorical variables even with large sample size. To alleviate the high rate of misclassification, a new strategy is embedded in the discriminant rule by introducing nonlinear principal component analysis (NPCA) into the classical location model (cLM), mainly to handle the large number of categorical variables. This new strategy is investigated on some simulation and real datasets through the estimation of misclassification rate using leave-one-out method. The results from numerical investigations manifest the feasibility of the proposed model as the misclassification rate is dramatically decreased compared to the cLM for all 18 different data settings. A practical application using real dataset demonstrates a significant improvement and obtains comparable result among the best methods that are compared. The overall findings reveal that the proposed model extended the applicability range of the location model as previously it was limited to only six categorical variables to achieve acceptable performance. This study proved that the proposed model with new discrimination procedure can be used as an alternative to the problems of mixed variables classification, primarily when facing with large categorical variables.

Keywords: Large categorical variables; leave-one-out method; location model; nonlinear principal component analysis; misclassification rate

ABSTRAK

Model lokasi yang dicadangkan pada masa lalu adalah satu peraturan diskriminan ramalan yang boleh mengelaskan cerapan baharu ke dalam salah satu daripada dua kumpulan yang telah ditetapkan berdasarkan campuran pemboleh ubah selanjar dan kategori. Keupayaan model lokasi untuk mendiskriminasi cerapan baharu dengan betul adalah amat bergantung kepada bilangan sel-sel multinomial yang dicipta melalui bilangan pemboleh ubah kategori. Penyelidikan ini menjalankan suatu kajian awal untuk menunjukkan model lokasi yang menggunakan anggaran kebolehjadian maksimum mempunyai kadar silap pengelasan yang tinggi sehingga 45% secara purata dalam berurusan dengan lebih daripada enam pemboleh ubah kategori bagi kesemua 36 data yang diuji. Model tersebut menunjukkan ramalan tidak tepat yang sangat tinggi kerana model ini berprestasi teruk bagi pemboleh ubah kategori besar walaupun dengan saiz sampel yang besar. Untuk mengurangkan kadar kesilapan pengelasan yang tinggi, satu strategi baharu telah diterapkan dalam peraturan diskriminan dengan memperkenalkan analisis komponen utama tak linear (NPCA) ke dalam model lokasi klasik (cLM), terutamanya untuk mengendalikan bilangan besar pemboleh ubah kategori. Strategi baharu ini dikaji pada beberapa set data simulasi dan sebenar melalui anggaran kadar silap pengelasan menggunakan kaedah leave-one-out. *Hasil daripada kajian berangka menampakkan kebolehlaksanaan model yang dicadangkan dengan kadar silap pengelasan menurun secara mendadak berbanding dengan cLM untuk kesemua 18 tetapan data yang berbeza. Aplikasi amali menggunakan set data sebenar menunjukkan penambahbaikan yang signifikan dan mendapat keputusan yang setanding dalam kalangan kaedah terbaik yang dibandingkan. Hasil kajian secara keseluruhan menunjukkan bahawa model yang dicadangkan memperluaskan rangkaian kebolehgunaan model lokasi kerana sebelum ini ia telah dihadkan kepada hanya enam pemboleh ubah kategori untuk mencapai prestasi yang boleh diterima. Kajian ini membuktikan bahawa model yang dicadangkan dengan prosedur diskriminasi yang baharu boleh digunakan sebagai alternatif kepada masalah klasifikasi pemboleh ubah campuran, terutamanya apabila berhadapan dengan pemboleh ubah kategori besar.*

Kata kunci: Analisis komponen utama tak linear; kadar silap pengelasan; kaedah leave-one-out*; model lokasi; pemboleh ubah kategori besar*

Introduction

The statistical treatment to analyse high dimensional multivariate data becomes a very significant methods for real life applications (Donoho 2000; Fan & Lv 2010). This situation has drawn attentions to design high dimensional discrimination for data consisting of continuous and categorical variables. For such purpose, this study considers the problem of mixed variables discrimination between two groups purposely to tackle large categorical variables in the classical location model.

The classical multivariate location model (cLM) proposed by Krzanowski (1975) is a predictive discriminant rule that can be used to assign new observations into one of the two predefined groups (Hamid & Mahat 2013). This cLM is developed through the utilization of both continuous and categorical variables (Hamid 2010). In particular, the continuous variables are utilized to estimate a set of parameters required in each multinomial cells created by the categorical variables. Based on the natural structure of cLM, all categorical variables need to be converted into binary structure in order to create segmentation called multinomial cells (Krzanowski 1993, 1983, 1975; Mahat et al. 2007). These multinomial cells play an important role in discriminating a new observation into group correctly (Asparoukhov & Krzanowski 2000; Krzanowski 1995). This is because the combination of the binary values gives rise to different multinomial cells which act as segmentation within group based on their mean values. This stressed that the ability of the cLM to discriminate new observations into the correct group is highly dependent on the number of multinomial cells created by the binary variables.

However, the main drawback of the cLM is the limitation of the number of binary variables in the construction of its predictive discriminant rule (Asparoukhov & Krzanowski 2000; Hamid & Mahat 2013; Krzanowski 1995, 1983; Mahat 2006). This is due to the structure of the location model itself as its multinomial cells grow exponentially with the binary variables following $s = 2^b$, where s is the number of multinomial cells created; and b is the number of binary considered in the study. Thus, the larger the binary variables the higher the multinomial cells will be created. Consequently, more chances that those multinomial cells will be empty as well as more estimated parameters will be bias hence lead to an unreliable model (Mahat et al. 2009).

A preliminary investigation has been conducted in this study to examine the classification performance of the cLM under various data conditions. These artificial data were assumed to have normal distribution with homogeneous covariance matrix across groups and cells. The variance of both binary and continuous variables is generated from an interval specified by the `rangeVar` function in R. Then, the covariance matrix is built through `rcorrmatrix` function in R. The data conditions were simulated such that they contain mixtures of continuous and binary variables with different sample sizes. There are four sample sizes are fixed as n=100, n=200, n=300 and n=400. Each set of sample sizes contains 10 continuous variables with various binary variables ranged from b=2 to b=10.

In Table 1, on average it can be seen that the cLM using maximum likelihood estimation shows misclassification rates greater than 40% in dealing with more than six binary variables. This result is in line with the study of Krzanowski (1975). These preliminary experiments showed that the misclassification rate increases when the percentage of empty cells increases. The percentage of empty cells has been recorded in the second row of Table 1. Their studies often limit the number of binary variables into some acceptance value i.e. at most six binary variables, so that the location model can be constructed with satisfactory performance. In this study, we focus on enlarges the sample size ranging from n=100 to n=400, however the misclassification rate is still found to be considerably high especially for large binary variables.

One common way to reduce the misclassification rate is by increasing the sample size to provide sufficient information in each multinomial cell. Alternatively, reducing the number of variables could be considered. However, Katz (2006) highlighted that increasing the size of sample is more desirable, but it is usually impossible to be carried out in any study. Besides, the sample size is usually limited in practise. Therefore, most researchers will try to reduce the number of variables involved in their studies. Li (2006) also stressed that reducing the dimension of the data helps to improve both recognition accuracy and efficiency. Keeping the dimensionality of measured variables as compact as possible is more desirable to obtain the most significant features that can describe important phenomenon of data and eliminate the redundant information (Young 2009).

Adoption of dimensional reduction such as variable selection or variable extraction can be beneficial to downsizing the variables (Zheng & Zhang 2008). Past studies has implemented different techniques of variable selection in the construction of the location model (Krzanowski 1995, 1983; Mahat et al. 2007). In order to use this approach, a subset of variables must be selected carefully without losing much information (Fan & Lv 2010). Otherwise, variable extraction can be an alternative method to obtain an extracted subset without abandoning some measured variables (Ramadevi & Usharaani 2013). Extracting significant variables is not only for the reason of computational time but also to improve the accuracy of the multivariate analysis (Ramadevi & Usharaani 2013). It has been shown when the techniques of variable extraction such as principal component analysis (PCA) and multiple corresponding analysis (MCA) were integrated with another context of location model using nonparametric smoothing estimation, positive improvement in accuracy follows (Hamid 2014, 2010; Hamid & Mahat 2013).

The purpose of this study was to highlight the possibility of the cLM in considering large number of categorical variables. In order to achieve this purpose, nonlinear principal component analysis (NPCA) is proposed to extract large categorical variables to allow for the

TABLE 1. The misclassification rate (ε) versus the percentage of empty cells (m_e) and computational time (t) of the classical location model for 10 continuous variables (c) with different number of binary variables (b) and sample size

Sample size = 100, c = 10									
	b = 2	b = 3	b = 4	b = 5	b = 6	b = 7	b = 8	b = 9	b = 10
ε (%)	3.00	20.00	29.00	40.00	41.00	42.00	46.00	46.00	54.00
m_e (%)	0.00	0.00	3.13	12.50	52.34	69.92	83.98	91.60	95.22
t (min)	0.058	0.090	0.139	0.240	0.630	0.816	1.550	3.139	7.084
Sample size = 200, c = 10									
	b = 2	b = 3	b = 4	b = 5	b = 6	b = 7	b = 8	b = 9	b = 10
ε (%)	1.00	6.00	13.50	28.50	36.00	41.00	47.50	51.50	48.50
m_e (%)	0.00	0.00	0.00	3.13	31.25	53.13	70.70	83.89	41.46
t (min)	0.109	0.180	0.311	0.569	0.949	1.848	3.798	7.422	16.622
Sample size = 300, c = 10									
	b = 2	b = 3	b = 4	b = 5	b = 6	b = 7	b = 8	b = 9	b = 10
ε (%)	2.00	6.30	12.30	24.30	28.30	40.70	42.00	44.00	43.00
m_e (%)	0.00	0.00	0.00	4.69	13.28	39.84	63.28	77.73	95.22
t (min)	0.159	0.241	0.428	0.811	1.555	3.161	5.821	11.663	21.043
Sample size = 400, c = 10									
	b = 2	b = 3	b = 4	b = 5	b = 6	b = 7	b = 8	b = 9	b = 10
ε (%)	1.30	2.00	6.80	13.80	28.50	30.50	43.30	45.50	47.30
m_e (%)	0.00	0.00	0.00	1.56	12.50	35.16	50.78	71.68	84.38
t (min)	0.264	0.536	0.640	1.226	2.175	4.305	8.434	17.050	35.542

construction of the LM using maximum likelihood based estimators.

PREDICTIVE DISCRIMINANT RULE OF THE CLASSICAL LOCATION MODEL

Suppose that the total sample size (n) is available from two groups. A set of n_1 observations belongs to group 1(π_1) and another set of n_2 observations belongs to group 2(π_2). Let a vector of b binary variables, $\mathbf{x}^{\mathrm{T}} = (x_1, x_2,\ldots, x_b)$ and a vector of c continuous variables, $\mathbf{y}^{\mathrm{T}} = (y_1, y_2,\ldots, y_c)$ are composed as an observed vector on each observation, $\mathbf{x}^{\mathrm{T}} = (x_1, x_2,\ldots, x_b)$. The b binary variables can be expressed as multinomial cells, $m = (m_1, m_2,\ldots, m_s)$ where $s = 2^b$. Before the expression of the binary variables, a general categorical variable with h levels is converted into $h - 1$ binary variables, then $\mathbf{x}$ is formulated with the number of multinomial cell by $m = 1 + \sum_q^b x_q 2^{q-1}$. Meanwhile, the c continuous variables are assumed to have multivariate normal distribution with mean, $\boldsymbol{\mu}_{im}$ in each multinomial cell of the two groups and a homogeneous covariance matrix, Σ across multinomial cells and groups. The probability of an observation in cell m of each group is denoted as ρ_{im} where i = 1, 2 and m = 1, 2, …, s. Thus, a new observation $\mathbf{z}^{\mathrm{t}} = (\mathbf{x}^{\mathrm{t}}, \mathbf{y}^{\mathrm{t}})$ is classified into π_1 using the following predictive discriminant rule if $\mathbf{x}$ falls in cell m and $\mathbf{y}$ satisfies

$$(\boldsymbol{\mu}_{1m} - \boldsymbol{\mu}_{2m})^T.\ \Sigma^{-1}\left\{\mathbf{y} - \frac{(\boldsymbol{\mu}_{1m} + \boldsymbol{\mu}_{2m}}{2}\right\} \geq \log\left(\frac{\rho_{2m}}{\rho_{1m}}\right) + \log(a), \quad (1)$$

otherwise, the new observation is classified to π_2. The constant a in (1) is dependent on the misclassifying costs and prior probabilities for the two groups. For simplicity, the costs as well as the probabilities are assumed equal in both groups, so $\log(a) = 0$. This predictive discriminant rule can be derived easily by replacing all parameters $\boldsymbol{\mu}_{im}$, Σ and ρ_{im} using maximum likelihood estimation from the data sample.

At first, the mean of each group of cell m needs to be estimated from a set of c continuous variables using

$$\hat{\boldsymbol{\mu}}_{im} = \sum_{j=1}^{n_{im}} \mathbf{y}_{jim} \big/ n_{im}, \quad (2)$$

where j = 1, 2, …, c; n_{im} is the number of observations in cell m of π_i; and $\mathbf{y}_{jim}$ is the vector of continuous variables of r^{th} observation in cell m of π_i.

Then, these estimated means are used to estimate a homogeneous covariance matrix through

$$\hat{\Sigma} = \sum_{i=1}^{2}\sum_{m=1}^{s}\sum_{j=1}^{n_{im}}\left(\mathbf{y}_{jim} - \hat{\boldsymbol{\mu}}_{im}\right)\left(\mathbf{y}_{jim} - \hat{\boldsymbol{\mu}}_{im}\right)^T \Big/ \left(n_1 + n_2 - s_1 - s_2\right), \quad (3)$$

where s_i is the number of non-empty cells in π_i.

Lastly, the cell probability can be estimated using

$$\hat{\rho}_{im} = n_{im} / n_i, \quad (4)$$

where n_i is the number of observations in π_i.

These three estimated parameters that are obtained from (2), (3) and (4), respectively, will be substituted into the (1) during the construction of the predictive discriminant rule of the cLM. However, it is only applicable for a limited number of categorical variables. This deficiency is manifested in the situation when many categorical variables are considered. Therefore, this study attempts to propose a new discrimination procedure based on the cLM for large categorical variables.

Design of Methods

As shown in Table 1, the estimation problems can arise and escalate as the number of b gets larger. Indeed, facing large categorical variables is a new challenge in the development of the location model. Analysis of all the measured variables might lead to data redundancy due to excessive variables included (Gervini & Rousson 2004). Furthermore, unnecessary large number of variables will burden the computational effort and require longer computational time as well (Ramadevi & Usharaani 2013). For these reasons, most high dimensional data is extracted to present a more compact information with a better visualization and accuracy (Mohd Aris et al. 2014). Variable extractions techniques have played an important role in the analysis of high dimensional data (Fan & Li 2006; Gervini & Rousson 2004). These techniques usually compute a system of b_q variables which are linear combinations of the original b variables that contribute most of the variation (Mohd Aris et al. 2014). These b_q variables are called principal components. Few variable extraction techniques have been demonstrated successfully in the development of some models. For example, the relevant application of principal component analysis (PCA) and multiple correspondence analysis (MCA) can be referred to Hamid and Mahat (2013) and Hamid (2014). These literatures highlighted that PCA is most appropriate to deal with large number of continuous variables while MCA is purposely designed to handle many categorical variables. The significant achievement of MCA in handling many categorical variables in Hamid (2014) has inspired this study to investigate the implementation of another variable extraction technique i.e. nonlinear principal component analysis into the location model.

Nonlinear Principal Component Analysis

Nonlinear principal component analysis (NPCA) is known as the extension to MCA and also the combination of PCA and optimal scaling of categorical variables (De Leeuw 2011). The goal of NPCA is similar to PCA but to reduce a large number of categorical variables to a smaller set that closely represents the original data with minimum loss of variation (Manisera et al. 2010). NPCA is unique as it utilizes the characteristics of PCA but it is able to play the similar role as MCA in handling categorical variables. The detail background, function description and case studies of NPCA have been described and explained thoroughly in the past literatures (De Leeuw 2011; De Leeuw & Mair 2009; Ferrari & Manzi 2010; Linting et al. 2007; Linting & Van der Kooij 2012; Manisera et al. 2010).

NPCA is a special case of homogeneity analysis that assumes and preserves the order of categories of the observed variables. Thus, it is specifically suitable for categorical variables. NPCA finds category quantifications that are optimal in the sense that the overall variance accounted for in the transformed variables is maximized (Linting & Van der Kooij 2012). In other words, as much as possible of the variance in the quantified variables is accounted for through the optimal quantification. Thus, the variables are transformed by assigning optimal scale values to the categories, providing in numeric-valued transformed variables.

In this study, the term categorical is referred as nominal variables that consist of unordered categories, so that this system is appropriate to be implemented in the construction of cLM. For example, there are two possible categories in gender, which are male and female. Such variables with unordered categories can be coded as zero for female and one for male. In fact, the optimal quantification can be any value as long as the objects of the same category received the same score on their quantified value. Therefore, NPCA's solution can be expressed through the optimal scaling process in quantification of the nominal variables.

In the analysis, NPCA provides eigenvalues as the overall summary measures that indicate the variance accounted for (VAF) by each component. VAF of a variable is defined as the sum of squared component loadings across component (Linting et al. 2007). This VAF acts as a good indicator to determine the number of components to be retained (Costa et al. 2013). Thus, the percentage of VAF has been used in this study as a stopping rule to justify relevant components in presence of categorical variables. This study uses 65% of VAF as a stopping rule following Solanas et al. (2011) and latest work by Long (2016). Table 2 presents the application procedures of NPCA.

Integration of the CLM and NPCA

With many advantages of NPCA found to accommodate the shortcoming of LM, therefore this study introduces a new strategy by integrating NPCA and cLM purposely to handle mixed variables discrimination for large categorical variables. This new strategy is performed based on the new discrimination procedure as summarized in Algorithm 1.

Simulation and Real Datasets

The proposed model is investigated in various conditions using some simulated and real datasets. The proposed model which is the integration of NPCA and cLM was

TABLE 2. The application procedures of NPCA

Procedures	Description
Data quantification	Perform the optimal quantification on the categorical data
Principal components	Construct the components in such a way that as much as possible of the variance in the quantified data is accounted for. The first constructed component explains the largest amount of VAF while the subsequent constructed component will be the second largest of the VAF
Stopping rule	Select the components to be retained based on the percentage of VAF with at least 65%

ALGORITHM 1. New proposed model via integration of cLM and NPCA

Step 1	Omit an observation k from the sample n as a test set, where $k = 1, 2, \ldots, n$.
Step 2	Perform NPCA on the binary variables from the remaining $n - 1$ observations to extract only significant components with VAF $\geq 65\%$.
Step 3	Fuse the extracted components in Step 2 together with the original continuous variables.
Step 4	Estimate $\boldsymbol{\mu}_{im}$, Σ and ρ_{im} using the fused data in Step 3.
Step 5	Construct the cLM rule using the parameters estimated in Step 4.
Step 6	Classify the group of a test set using the rule constructed in Step 5.
Step 7	Record the misclassifying group of the object in Step 6 as $\varepsilon_k = 1$, otherwise $\varepsilon_k = 0$.
Step 8	Repeat Step 1 to Step 7 for all observations in turn.
Step 9	Compute the average of the misclassification rate using $\sum_{k=1}^{m} \varepsilon_k \big/ n$.

evaluated through the estimation of the misclassification rate using leave-one-out (LOO) method. This procedure omits an observation from the sample as a test set for validation purpose and utilizes the remaining observations as a training set to construct the proposed model. Meanwhile, the misclassification rate was calculated by taking the difference between the actual group and the predicted group of the test set. Such procedures were repeated until all observations had been omitted in turns. To initialize the simulated data, all the key factors such as n, b and c were set differently to investigate the proposed model from a wide range of conditions as possible within reasonable practical scopes. Sample sizes are set at n = 100 and n = 200 where each set contains 5, 10 and 15 binary variables. The b = 10 and b = 15 can be considered as very large number of binary variables in the context of location model due to the structure of the location model itself, whereby these binary sizes will produce specifically 1024 and 32768 cells per group, respectively.

Meanwhile, the number of continuous variables (c) was set to 5, 10 and 15 to create different conditions such as $b < c$, $b = c$ and $b > c$ as displayed in Table 3. This study generates variance of $c + b$ variables ($\sigma_1, \ldots, \sigma_{(c+b)}$) from an interval specified by the `rangeVar` function in R. Next, we find a correlation matrix (**R**) using the build-in function called `rcorrmatrix`. Thus, the covariance matrix is generated through

$$\boldsymbol{\Sigma} = dd \times \mathbf{R} \times dd, \tag{5}$$

where dd is the diagonal elements of ($\sigma_1, \ldots, \sigma_{(c+b)}$).

Altogether this study has 18 different simulated datasets representing different conditions for evaluating the performance of the proposed model. However, for the purpose of this study, the existence of correlation among the variables is considered do not exist.

In order to investigate on the possible extent of the proposed model in practical applications, a real dataset of full breast cancer from King's college Hospital, London was used. This data consists of 137 patients having breast cancer and was divided into two groups i.e. 78 are benign (π_1) and 59 are malignant (π_2). This full breast cancer data contains 15 variables comprising of 2 continuous variables, 4 nominal variables with three states each, 6 ordinal variables with eleven states each and 3 binary variables. All the ordinal variables are treated as continuous variables while all the nominal variables are transformed into binary variables according to the past studies (Hamid 2014; Krzanowski 1975; Mahat et al. 2007). This pre-processing gives a new dimension with eight continuous and eleven binary variables of a full breast cancer data.

This full breast cancer data is considered mimics to the simulated data that are generated in this study. There are 11 binary variables, which is in the range of b = 10 and b = 15 as in the simulation dataset. In addition, the sample size of this data is 137 which falls between n = 100 and n = 200 in the simulation study.

The discrimination performance of the proposed model was then compared with some of the existing discriminant methods for this real dataset.

TABLE 3. Various datasets to evaluate and compare the performance of the proposed model and cLM

Number of continuous and binary variables	$n = 100$			$n = 200$		
	ε (%)	m_e(%)	t	ε (%)	m_e(%)	t
For $c = 5$						
$b = 5$	Dataset 1, b = c			Dataset 10, b = c		
$b = 10$	Dataset 2, b > c			Dataset 11, b > c		
$b = 15$	Dataset 3, b > c			Dataset 12, b > c		
For $c = 10$						
$b = 5$	Dataset 4, b < c			Dataset 13, b < c		
$b = 10$	Dataset 5, b = c			Dataset 14, b = c		
$b = 15$	Dataset 6, b > c			Dataset 15, b > c		
For $c = 15$						
$b = 5$	Dataset 7, b < c			Dataset 16, b < c		
$b = 10$	Dataset 8, b < c			Dataset 17, b < c		
$b = 15$	Dataset 9, b = c			Dataset 18, b = c		

RESULTS AND DISCUSSION

RESULTS FROM SIMULATION STUDY

This section gives the findings on the evaluations conducted on the proposed model and cLM. The study begins by looking at the issue of empty cells occurred and its relation to the discrimination accuracy. The investigation continues with the performance of these two models based on the number of binary variables and the size of sample through the evaluation of the misclassification rates. Finally, this study compares the computational time needed to complete their tasks for each discrimination process.

It is a fact that the number of multinomial cells (s) grows exponentially according to the number of binary variables (b) due to the structure of the location model itself following $s = 2^b$. For example, 5 binary variables will create 32 cells per group while 15 binary variables will create 32,768 cells per group. Therefore, these cells will have high chances to become empty if involved with large categorical variables, particularly when the samples sizes are small. In this situation, the proposed model has advantage over the cLM as shown in Figure 1. This graph demonstrates that the percentage of empty cells occurred in cLM is much higher than the proposed model for all datasets investigated. The percentage of empty cells occurred in cLM achieved more than 90% for both $b = 10$ and $b = 15$ (Table 3). In line with this, the results displayed in Table 3 proves that the misclassification rates are high, ranging from 41.5% to 51.5%, for those large percentage of empty cells. In contrast, the proposed model utilizes the strengths of NPCA to extract the large binary variables into a smaller number of components with maximum variance explained. The outputs from all 18 datasets verified that NPCA manages to intensely reduce the percentage of empty cells i.e. more than half, on average. Due to the ability of NPCA, thus the performance of the proposed model shows significant improvement for all datasets as shown in Table 3.

Handling large binary variables directly in the cLM is very burdensome due to excessive number of multinomial

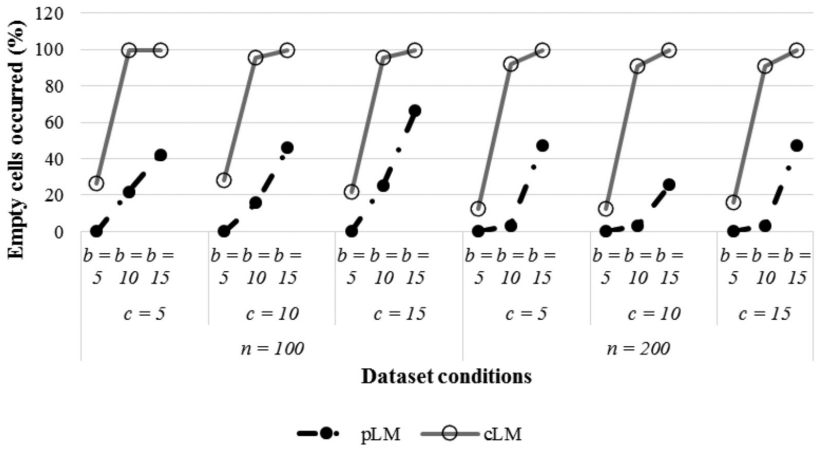

FIGURE 1. The percentage of empty cells occurred for the proposed model versus cLM

cells that will be created and high possibility that these cells are empty. Consequently, the performance of cLM is poor in dealing with large binary variables for both $n = 100$ and $n = 200$. Thus, this study attempts to downsize the large categorical variables using NPCA before the construction of cLM. The results highlighted that the extraction of binary variables is effective in improving the accuracy of discrimination performance as the misclassification rate is decreased for all conditions of data examined. This demonstrates that the dimensionality reduction is beneficial for discrimination purposes, mainly when facing with a large number of measured variables. As can be seen from Table 4, the percentage of empty cells is much lower with NPCA technique and further improves the performance of the proposed model. This study further displays the performance of the proposed model and cLM simultaneously in Figure 2. The results clearly show the accuracy of the proposed model is greatly better as compared to the cLM for both $n = 100$ and $n = 200$ as well as for different combinations of b and c.

In terms of time required to complete the task, as can be seen from Table 3, the computational time of the proposed model is much shorter than the cLM for large binary variables i.e. $b = 10$ and $b = 15$. The computational time of cLM needs more than 10 h to 2 days for 15 binary variables. With the help of NPCA, the computational time to run 15 binary variables is reduced dramatically from days to few min. In addition to long computational hours of the cLM, its discrimination accuracy is only guaranteed on average of 50% for a correct prediction. On the other hand, for those cases with 5 binary variables, cLM is able to obtain a quick feedback but its misclassification rate is significantly higher than the proposed model.

In summary, these simulation results describe the effects of empty cells and the number of binary variables as well as the sample sizes towards the misclassification rate of a predictive discriminant rule. In simple words, there are two significant relationships identified from this study. First, the smaller the percentage of empty cells, the lower will be the misclassification rate of a discriminant rule. The dimensionality reduction plays an important role as data pre-processing especially for large dimensional multivariate analysis as it manages to reduce the percentage of empty cells of the location model. Second, the larger the sample size, the misclassification rate of a discriminant rule is slightly lower. The large sample size provides adequate information for multinomial cells which may improve the discrimination performance.

Results From Real Empirical Study

There are seven existing discrimination methods used to investigate and compare the performance of the proposed model as presented in Table 5. We take the results of other existing discrimination methods from past study by Mahat (2006) which represent three groups of statistical approaches i.e. parametric, semi-parametric and non-parametric approaches such as linear discriminant analysis, quadratic discriminant analysis, logistic discrimination, regression model, tree classification as well as location model using nonparametric smoothing estimation. This study further compares and validates the performance of the proposed model among these discrimination methods plus with the classical LM.

The outcomes in Table 5 show that the integration of non-parametric smoothing of LM and variables extraction techniques (PCA+MCA+LM and 2PCA+LM) performs the best. Then, the third best discriminant rule goes to logistic discrimination which includes all the variables in the model development, followed by the proposed model (NPCA+cLM). This finding tells us that discriminant rules with variable extractions including the proposed one are among the winner and followed by the discriminant rules that include all variables except quadratic discrimination and cLM.

The difference between the proposed model i.e. integration of cLM with variables extraction technique (NPCA+cLM) and cLM is obvious where the former shows great improvement from the latter. This result implies that the implementation of NPCA is able to work well

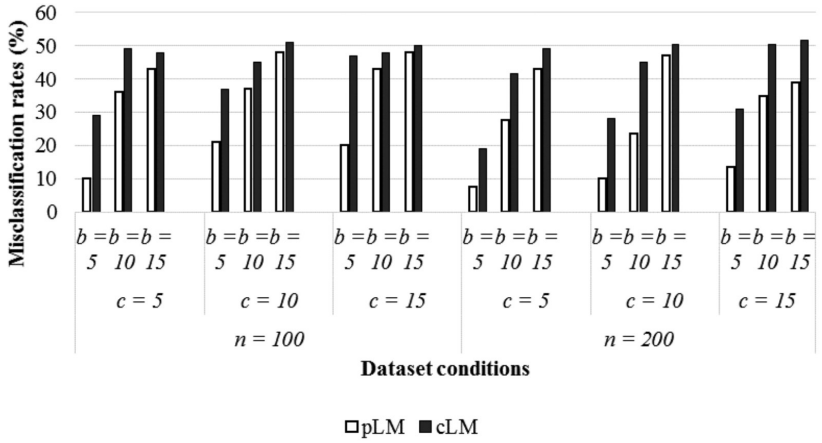

FIGURE 2. The misclassification rates of the proposed model versus cLM

TABLE 4. The performance of the proposed model (A) and cLM (B) for all data conditions

	$n_1 = n_2 = 50, c = 5$					
	ε (%)		m_e(%)		t	
	A	B	A	B	A	B
$b = 5$	10.00	29.00	0.00	26.56	33.16 s	8.53 s
$b = 10$	36.00	49.00	21.88	99.90	2.36 min	3.99 min
$b = 15$	43.00	48.00	42.19	99.85	6.80 min	10.36 h
	$n_1 = n_2 = 50, c = 10$					
	ε (%)		m_e(%)		t	
	A	B	A	B	A	B
$b = 5$	21.00	37.00	0.00	28.13	45.36 s	21.13 s
$b = 10$	37.00	45.00	15.63	95.46	2.60 min	12.26 min
$b = 15$	48.00	51.00	46.09	99.85	4.12 min	17.70 h
	$n_1 = n_2 = 50, c = 15$					
	ε (%)		m_e(%)		t	
	A	B	A	B	A	B
$b = 5$	20.00	47.00	0.00	21.88	35.29 s	13.44 s
$b = 10$	43.00	48.00	25.00	95.41	2.43 min	6.21 min
$b = 15$	48.00	50.00	66.41	99.85	5.34 min	1.09 day
	$n_1 = n_2 = 100, c = 5$					
	ε (%)		m_e(%)		t	
	A	B	A	B	A	B
$b = 5$	7.50	19.00	0.00	12.50	1.19 min	20.75 s
$b = 10$	27.50	41.50	3.13	91.80	6.07 min	9.80 min
$b = 15$	43.00	49.00	47.27	99.70	15.67 min	21.30 h
	$n_1 = n_2 = 100, c = 10$					
	ε (%)		m_e(%)		t	
	A	B	A	B	A	B
$b = 5$	8.00	28.00	6.25	12.50	1.57 min	23.46 s
$b = 10$	23.50	45.00	3.13	91.11	7.09 min	12.24 min
$b = 15$	47.00	50.50	25.78	99.70	16.78 min	1.73 day
	$n_1 = n_2 = 100, c = 15$					
	ε (%)		m_e(%)		t	
	A	B	A	B	A	B
$b = 5$	13.50	31.00	0.00	15.63	1.78 min	27.31 s
$b = 10$	35.00	50.50	3.13	91.00	8.39 min	13.48 min
$b = 15$	39.00	51.50	47.26	99.70	15.11 min	1.99 day

with large categorical variables and greatly improve the discrimination performance of cLM. This finding also tells us that NPCA is succeeds in extracting only the important components that should be included in the final model developed. Such result gives the impression that NPCA may handle the process of extraction on the categorical variables.

The adaption of NPCA manages to increase the ranking of cLM from worst to top four in performance. In other words, the proposed model obtains comparative outcome among the best methods (nonparametric smoothing of LM with variable extractions and logistic discrimination) and performs better than the other six strategies. Meanwhile, the discriminant rules that include some of the variables, i.e. nonparametric smoothing of LM with variable selections and tree, show bad performance. This implies that most of the variables contribute to discriminate the benign and malignant patients.

TABLE 5. The performance ranking of the proposed model versus other existing discrimination methods for full breast cancer data

Discrimination method	Integrated strategy	Misclassification rate	Ranking
Classical location model	All variables included (cLM)	0.3942	12
	NPCA for binary variables (NPCA+cLM)	0.2920	4
Linear discriminant analysis	All variables included	0.2920	4
Quadratic discriminant analysis	All variables included	0.4453	13
Logistic discrimination	All variables included	0.2847	3
Regression model	Forward selection	0.3139	8
	Backward selection	0.2920	4
	Stepwise selection	0.2920	4
Tree classification	Auto-termination	0.3139	8
Nonparametric smoothing of location model	Forward selection	0.3139	8
	Backward selection	0.3139	8
	PCA for both continuous and binary variables (2PCA+LM)	0.2774	2
	PCA for continuous variables and MCA for binary variables (PCA+MCA+LM)	0.2336	1

From all the findings obtained, it illustrates that the application of dimensional reduction before the construction of a predictive discriminant rule is important and good for a large number of measured variables. However, the smoothing technique required large space and time to be constructed as it needs a weighted value for each iteration in the leave-one-out process. Thus, it can be concluded that the proposed model is another tool that can be used to obtain quick results with comparable performance and sometimes better, even with a large number of categorical variables.

CONCLUSION

This study examines a new strategy from the integration of NPCA with large categorical variables into the cLM. This new strategy applies NPCA in the classical location model for handling the issue of many empty cells. The results from simulation and real datasets have proved that the proposed model based on this new strategy can be an alternative to other discrimination methods, mainly when involved with large categorical variables. The methods proposed is a systematic procedure to extract meaningful categorical variables before the construction of the location model, which has been demonstrated in enhancing the discrimination performance. This study finding showed that the implementation of variable extraction could improve both the discrimination accuracy and the computational efficiency. As a conclusion, the investigations from simulation and real datasets provide additional and useful knowledge regarding mixed variables discriminant analysis and variable extraction technique on a large categorical variables in particular.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Utara Malaysia for the financial support.

REFERENCES

Asparoukhov, O. & Krzanowski, W.J. 2000. Non-parametric smoothing of the location model in mixed variable discrimination. *Statistics and Computing* 10: 289-297.

Costa, P.S., Santos, N.C., Cunha, P., Cotter, J. & Sousa, N. 2013. The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. *Journal of Aging Research* 2013: Article ID. 302163. doi:10.1155/2013/302163.

De Leeuw, J. 2011. *Nonlinear Principal Component Analysis and Related Techniques*. UCLA: Department of Statistics. https://escholarship.org/uc/item/7bt7j6nk.

De Leeuw, J. & Mair, P. 2009. Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software* 31(4): 1-21. http://www.jstatsoft.org/.

Donoho, D.L. 2000. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*. pp. 1-33. http://mlo.cs.man.ac.uk/resources/Curses.pdf.

Fan, J. & Li, R. 2006. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Feature Selection in Knowledge Discovery*. pp. 1-27. doi:10.4171/022-3/31.

Fan, J. & Lv, J. 2010. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20(1): 101-148. doi:10.1063/1.3520482.

Ferrari, P.A. & Manzi, G. 2010. Nonlinear principal component analysis as a tool for the evaluation of customer satisfaction. *Quality Technology and Quantitative Management* 7(2): 117-132. http://air.unimi.it/handle/2434/141402\nhttp://web2.

cc.nctu.edu.tw/~qtqm/qtqmpapers/2010V7N2/2010V7N2_F2.pdf.

Gervini, D. & Rousson, V. 2004. Criteria for evaluating dimension-reducing components for multivariate data. *The American Statistician* 58(1): 72-76. doi:10.1198/0003130042863.

Gupta, V. 2013. *Exploring Data Generated by Pocket Devices*. London. http://files.howtolivewiki.com/SMART_CITIES/The_Smart_City.To_Whos_Advantage.Pocket_Devices_and_Data_Trails.Vinay_Gupta.pdf.

Hamid, H. 2010. A new approach for classifying large number of mixed variables. *International Scholarly and Scientific Research and Innovation* 4(10): 120-125. doi:14621.

Hamid, H. 2014. Integrated smoothed location model and data reduction approaches for multi variables classification. Doctoral Dissertation. Universiti Utara Malaysia, Malaysia (Unpublished).

Hamid, H. & Mahat, N.I. 2013. Using principal component analysis to extract mixed variables for smoothed location model. *Far East Journal of Mathematical Sciences (FJMS)* 80(1): 33-54.

Katz, M.H. 2011. *Multivariate Analysis: A Practical Guide for Clinicians and Public Health Researchers*. Cambridge: Cambridge University Press.

Krzanowski, W.J. 1995. Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. *Computational Statistics & Data Analysis* 19: 419-431. doi:10.1016/0167-9473(94)00011-7.

Krzanowski, W.J. 1993. The location model for mixtures of categorical and continuous variables. *Journal of Classification* 10(1): 25-49. doi:10.1007/BF02638452.

Krzanowski, W.J. 1983. Stepwise location model choice in mixed-variable discrimination. *Journal of the Royal Statistical Society. Series C (Applied Statisitcs)* 32(3): 260-266.

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

Li, Q. 2006. An integrated framework of feature selection and extraction for appearance-based recognition. Doctoral Dissertation. University of Delaware Newark, DE, USA (Unpublished).

Linting, M., Meulman, J.J., Groenen, P.J.F. & Van der Kooij, A.J. 2007. Nonlinear principal components analysis: Introduction and application. *Psychological Methods* 12(3): 336-358. doi:10.1037/1082-989X.12.3.336.

Linting, M. & Van der Kooij, A.J. 2012. Nonlinear principal components analysis with CATPCA: A tutorial. *Journal of Personality Assessment* 94(1): 12-25. doi:10.1080/00223891.2011.627965.

Long, M.M. 2016. Binary variable extraction using nonlinear principal component analysis in classical location model. Master Dissertation. Universiti Utara Malaysia, Malaysia (Unpublished).

Mahat, N.I. 2006. Some investigations in discriminant analysis with mixed variables. Doctoral Dissertation. University of Exeter, London, UK (Unpublished).

Mahat, N.I., Krzanowski, W.J. & Hernandez, A. 2009. Strategies for non-parametric smoothing of the location model in mixed-variable discriminant analysis. *Modern Applied Science* 3(1): 151-163.

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. doi:10.1007/s11634-007-0009-9.

Manisera, M., A.J. Van der Kooij, & Dusseldorp, E. 2010. Identifying the component structure of job satisfaction by nonlinear principal components analysis. *Quality Technology and Quantitative Management* 7: 97-115. http://elisedusseldorp.nl/pdf/Manisera_QTQM2010.pdf.

Mohd Aris, Khairul Dahri, Faizal Mustapha, Mohd Sapuan Salit & Dayang Laila Abang Abdul Majid. 2014. Condition structural index using principal component analysis for undamaged, damage and repair conditions of carbon fiber-reinforced plastic laminate. *Journal of Intelligent Material Systems and Structures* 25(5): 575-584. doi:10.1177/1045389X13494932.

Ramadevi, G.N. & Usharaani, K. 2013. Study on dimensionality reduction techniques and applications. *Publications of Problems & Application in Engineering Research* 4(1): 134-140.

Russom, P. 2013. *Managing Big Data*. TWDI Best Practices Report. Washington: twdi.org.

Solanas, A., Manolov, R., Leiva, D. & Richard, M.M. 2011. Retaining principal components for discrete variables. *Anuario de Psicologia* 41(1-3): 33-50.

Vlachonikolis, I.G. & Marriott, F.H.C. 1982. Discrimination with mixed binary and continuous data. *Applied Statistics* 31(1): 23-31.

Young, P.D. 2009. Dimension reduction and missing data in statistical discrimination. Doctoral Dissertation. USA Baylor University (Unpublished).

Zheng, H. & Zhang, Y. 2008. Feature selection for high-dimensional data in astronomy. *Advances in Space Research* 41(12): 1960-1964. doi:10.1016/j.asr.2007.08.033.

Statistics Department, School of Quantitative Sciences
Universiti Utara Malaysia College of Arts and Sciences
06010 UUM Sintok, Kedah Darul Aman
Malaysia

*Corresponding author; email: hashibah@uum.edu.my

Received: 16 October 2015
Accepted: 28 November 2016