Ann. Data. Sci. (2015) 2(4):489–504
DOI 10.1007/s40745-015-0061-9

An Efficient Variable Selection Method for Predictive Discriminant Analysis

A. Iduseri[1] · **J. E. Osemwenkhae**[1]

Received: 24 September 2015 / Revised: 8 December 2015 / Accepted: 9 December 2015 /
Published online: 19 December 2015

Abstract Seeking a subset of relevant predictor variables for use in predictive model construction in order to simplify the model, obtain shorter training time, as well as enhance generalization by reducing overfitting is a common preprocessing step prior to training a predictive model. In predictive discriminant analysis, the use of classic variable selection methods as a preprocessing step, may lead to "good" overall correct classification within the confusion matrix. However, in most cases, the obtained best subset of predictor variables are not superior (both in terms of the number and combination of the predictor variables, as well as the hit rate obtained when used as training sample) to all other subsets from the same historical sample. Hence the obtained maximum hit rate of the obtained predictive discriminant function is often not optimal even for the training sample that gave birth to it. This paper proposes an efficient variable selection method for obtaining a subset of predictors that will be superior to all other subsets from the same historical sample. In application to real life datasets, the obtained predictive function using our proposed method achieved an actual hit rate that was essentially equal to that of the all-possible-subset method, with a significantly less computational expense.

Keywords Variable selection · Predictive discriminant analysis · Actual hit rate · Superior subset

✉ A. Iduseri
augustine.iduseri@uniben.edu

[1] Department of Mathematics, Faculty of Physical Sciences, University of Benin, P.M.B. 1154, Benin City 300001, Edo State, Nigeria

1 Introduction

Predictive discriminant analysis (PDA) is used in many situations where prior designation of groups exists: such as migration/non-migration status, employed/unemployed, which customers are likely to buy a product or not buy, whether a person is a credit risk or not, e.t.c. The aim of PDA is to classify an observation, or several observations, into these known groups. To reduce modeling bias, it is common to start with more predictor variables than are actually believed to be required. This may necessitate a substantial search to determine which predictor variables are most important to include in the prediction model. Researchers often gather data on features or predictor variables, which they believe are good discriminators. This may well be the case, for example, when researchers conduct a preliminary investigation trying to discover useful predictor variables. Thus, they might ask themselves questions such as (1) are all predictor variables really necessary for effective discrimination and (2) which predictor variables are the best predictors? This will in no doubt help determine a rule that will yield a high degree of classification precision as well as predictive accuracy" [1].

In many applications, only a subset of the predictor variables in PDA contain any group membership information, and including predictor variables which have no group information increases the complexity of the analysis, potentially degrading the classification performance [2] . There can be many reasons for selecting only a subset of the predictor variables instead of the whole set of candidate variables [3]: (1) It is cheaper to measure only a reduced set of predictor variables, (2) Prediction accuracy may be improved through exclusion of redundant and irrelevant predictor variables, (3) The predictor to be built is usually simpler and potentially faster when fewer input predictor variables are used and (4) Knowing which predictor variables are relevant can give insight into the nature of the prediction problem and allows a better understanding of the final classification model. Therefore, there is a need for including predictor variable selection as part of any PDA procedure.

Research in variable selection started in the early 1960s [4]. Over the past four decades, extensive research into feature selection has been conducted. Much of the work is related to medicine and biology [5–8]. Besides the classical methods such as the stepwise and all possible subset methods, a range of other approaches to variable or subset selection in classification context include the genetic search algorithms wrapped around Fisher discriminant analysis by [9]. We also have a number of different search algorithms (proposed as alternatives to backward/forward/stepwise search) wrapped around different discriminant functions compared by [10], variable selection for kernel Fisher discriminant analysis [11], DALASS approach of [12] and sequential stepwise analysis method [13]. A good review of methods involving support vector machines (SVM_S) along side with a proposed criterion for exhaustive variable or subset selection is given by [14]. Another approach to variable selection is the shrinkage method. Notable variants off the shrinkage method are the least absolute selection and shrinkage operator (LASSO) [15] and ridge regression (RR) [16]. As with RR, the LASSO shrinks the coefficient estimates towards zero. As a result, models generated from the LASSO are generally much easier to interpret than those produced by RR. A third approach to variable selection is the dimension reduction method. A popular approach

of this method which is used for deriving a low-dimensional set of features from a large set of variables is the principal components analysis (PCA) [16,17]. Lastly we have some recent representative methods which are based on dictionary learning (DL) for classification. The representative approach can conveniently be categorized as Track I and Track II [18]. Track I includes Meta-face learning [19] and DL with structured incoherence [20], while Track II includes supervised DL [21], discriminative K-SVD [22], label consistence K-SVD [23] and Fisher discrimination DL [24].

The classical variable selection strategy such as stepwise methods hold out the promise of assisting researchers with such important task as variable selection prior to training of the PDF in order to maximize hit rate. However, the promise of maximizing the hit rate is almost always unfulfilled [1,25,26]. The all-possible subset method which is a notable classic variable selection strategy developed to address the problems inherent with the stepwise methods also suffers from computational limitations [27]. Also, other approaches to variable selection in a classification context majorly suffer from computational limitations [28]. Most of the variable selection methods that have been developed to assist the researcher in deciding which predictors to prune and the ones to keep, simply involve selection of best subset of variables under some criterion, such as the smallest R^2_{adj}, mean square error (MSE), Mallows' C_K, Bayes' information criterion (BIC), Akaike's information criterion (AIC), e.tc. These methods search for subset of variables that can carry out this classification task in an optimum way, with the hope of reducing computational time such that the predictive function solution (or hit rate) is optimal. In the context of PDA, best subset is the subset with highest hit rate. However, a better variable selection method may be obtained if its "best" subset with the highest hit rate is superior to all other subsets from the same historical sample.

This work proposes an efficient variable selection method for obtaining a best subset of predictor variables that is superior (in terms of having a higher hit rate when compared to any other subsets with the same number of predictor variables) to all other subsets of predictors from the same historical sample with less computational expense.

2 Classical Variable Selection Methods Used in Predictive discriminant Analysis

The task of improving the performance (or maximizing hit rate) of the predictive discriminant function (PDF) usually begins with the researcher making choices about the variables that will be involved in the analysis. This may necessitate a substantial search to determine which variables are most important to include in the predictive model. Although there are many references in the literature regarding selecting variables for their use in classification, there are very few key references on the selection of variables for their use in PDA. In PDA, the most commonly used variable selection strategy is either the stepwise methods [29,30] or the all possible subset method [26,31] proposed to address problems inherent with stepwise methods [26]. These classical variable selection procedures are readily available via common statistical computer packages/programmes such as SPSS, SAS, R, e.t.c. We give a brief review of these two notable classic variable selection methods.

2.1 Stepwise Methods

Several variants of stepwise methods that are readily available via common statistical computer packages include the forward variable selection, backward variable selection, and stepwise variable selection. The Forward method (or forward selection) begins by selecting the most discriminating variable according to some criterion. It continues by selecting the second most discriminating variable and so on. The algorithm stops when none of the non-selected variables discriminates in a significant way or when the increase in the coefficient of determination (R^2) is no longer statistically significant. The backward method (or backward elimination) works in the opposite way. When using the backward selection process, all the variables are initially included in the model. As the analysis progresses, any predictor variable that does not contribute to the model is deleted. The algorithm stops when all the remaining variables discriminate significantly.

The stepwise variable selection uses a combination of the two previous algorithms: at each step variables are introduced or eliminated depending on how significant their discriminating capacity is. It also allows for the possibility of changing decisions taken in previous steps, by eliminating from the selected set a variable introduced in a previous step of the algorithm or by selecting a previously eliminated variable. The basic difference between the forward selection process and the stepwise process is that the stepwise process, before entering a new predictor variable, checks to see if all the predictors already in the model remain significant. Thus, if a previously selected predictor is no longer useful, the procedure will drop that predictor variable. On the hand, using the forward selection method, once a predictor enters the model, it remains there. Stepwise procedures need a mechanism for controlling the entry or removal of predictor variables from the PDF. Notable methods for controlling the entry or removal of predictor variables from the PDF are (a) lambda, (b) mahalanobis distance, (c) smallest F ratio, (d) Rao's V, and (e) sum of unexplained variance. Klecka [32] pointed out that "the end result will often be the same regardless of the criterion used, but it is not always the case". However, for the stepwise methods, if the independents are correlated (or have shared discriminating power), an important predictor may not be selected, because it's unique contributions are not as great as those of other variables. Since the stepwise procedure is a logical way to seek the best combination of predictor variables, it cannot guarantee that the end product is indeed superior to all others [33].

2.2 All-Possible Subset Approach

The all-possible subset approach, as the name implies, analyzes the data one-predictor at a time, two-predictor at a time, and so on. Thus, as the number of predictors increases, so does the number of analyses. In fact, for p predictors, a total of $2^p - 1$ predictor subsets would be assessed. For example, when there are four predictors there would be $2^4 - 1 = 15$ predictor subsets to be assessed. Hence, the all-possible subset method becomes computationally infeasible for values of predictors greater than around 40; even with extremely fast modern computers [27]. Also, If there are two (or more) subsets of a given size that yields the same hit rates, the subset of choice is either based

on researcher's judgment or a second information criterion is used. However, the all possible subsets approach has remained a popular alternative to stepwise procedure. For this reason, the all possible subset procedure will be used for the purpose of comparative analysis.

3 Developing the Predictive Discriminant Function for Future Use

In PDF, having obtained a best subset of predictor variables using any of the notable variable selection methods the next step is to train the PDF. However, linear model users are often disappointed when the model that predicts group membership well for the original data set becomes at best marginal when applied to fresh data drawn from the same population or historical sample. This is often the case because the predictive models do capitalize on chance and therefore lead to situations where the function may predict group membership of the initial data set far better than any other data set or sample that could be drawn. Clearly, testing the PDF on the data that gave birth to it is almost certain to overestimate performance [26]. For the optimizing process that chose it from among many possible PDF_S will have made the greatest use possible of any and all idiosyncracies of those particular data.

Thus using the classical variable selection methods or any other notable approach to obtain a best subset of predictor variables, that is superior to all other subsets of predictors from the same historical sample in order to obtain an optimal PDF solution is not a guarantee that it will be able to generalize to new measurements. That is why we sometimes say that optimization capitalizes on chance [34]. Optimization based on chance creates a degree of fit, but in the case of the predictive analysis, this fit may be upward biased and not representative of real world situations [35]. The general solution is to evaluate the PDF by testing it on new data sets distinct from the training sample that gave birth to it using validation and/or cross-validation (CV) procedures. Generally, CV is a way to predict the fit of a model to a hypothetical validation set when an explicit validation set (i.e., a validation set that is completely drawn separately but from the same population) is not available. In PDA, CV is mostly used in estimation of actual hit rate or true error rate [36,37]. Notable classical variants of CV procedures due to different splitting strategies are the Leave-one-out cross validation [38–40] and K-fold cross validation [40] methods. We also have the percentage-N-fold cross validation rule [41] proposed to address the choice of k.

3.1 Leave-One-Out Cross-Validation Method

Leave-one-out CV is the most classical exhaustive CV procedure originally designed to estimate the actual hit rate, $P^{(a)}$. It involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This process is repeated N times such that each observation in the historical sample is used once as the validation data, and the proportions of deleted units (test samples) correctly classified are used to estimate the hit-rate. If d_j is defined as

$$d_j = \begin{cases} 1 & \text{if } \hat{Z}_j^{-j} = Z_j \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{Z}_j^{-j}$ is the predicted response for the jth observation computed with the jth observation removed from the historical sample, Z_j is the value of jth observation in the historical sample. Mathematically, the LOOCV estimate of the actual hit-rate, $P^{(a)}$ is given by

$$\hat{P}_{LOOCV}^{(a)} = \frac{1}{N}\left(\sum_{j=1}^{N} d_j\right) \times 100 \tag{3.1}$$

where N is the total number of cases over all groups (or size of historical sample)

3.2 K-Fold Cross-Validation Method

K-fold cross-validation (KFCV) was introduced as an alternative to the computationally expensive LOOCV [42]. Breiman and Spector [43] found 10-fold and 5-fold CV to work better than LOOCV. Generally, in literature a value of 10 for K is popular for estimating the error rate or hit rate. If we denote $P^{(a)}$ as the hit rate for each of the K sub-samples, and let

$$d_j = \begin{cases} 1 & \text{if } \hat{Z}_j^{-k(j)} = Z_j \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{Z}_j^{-k(j)}$ is the predicted response for the jth observation computed with the k(j)th part of the data removed, Z_j is the value of jth observation in the historical sample, $k\,(j)$ is the fold containing observation j. Then the hit rate for the K sub-sample is define as

$$P_k^{(a)} = \frac{1}{n_v}\left(\sum_{j=1}^{n_v} d_j\right) \times 100 \tag{3.2a}$$

and n_v is the number of cases in the validation sample. Therefore, the KFCV estimate of the actual hit rate is given as

$$\hat{P}_{KFCV}^{(a)} = \frac{1}{K}\sum_{k=1}^{K} P_k^{(a)} \tag{3.2b}$$

3.3 Percentage-N-Fold Cross-Validation Rule

The percentage-N-fold CV rule can be seen as an alternative to the K-fold CV method. It has been effectively used in estimation of actual hit rate or the true error rate in discriminant analysis [44].

The outline of the percentage–N-fold CV (NFCV$_{-P}$) procedure is given as follows:

Step1: Obtain a training set, $I^{(t)}$ as a percentage of the historical sample, D_N
Step2: For each training sample, $D_N^{(t)}$ obtained in step1, compute $Z = \eta\left(D_N^{(t)}\right)$ and obtain it's $P^{(a)}$ on the Historical sample, D_N
Step3: Repeat steps 1-2 using percentage values of 60, 70, 80 and 90 respectively.
If we denote $\hat{P}_{(n)}^{(a)}$ as the estimate of $P^{(a)}$ for each of the N sub-samples, and let

$$d_j = \begin{cases} 1 & \text{if } \hat{Z}_j^{-n(j)} = Z_j \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{Z}_j^{-n(j)}$ is the predicted response for the jth observation computed with the (j)th part of the data removed from the historical sample, Z_j is the value of jth observation in the historical sample, n(j) is the fold containing observation j. Then the estimate, $\hat{P}_{(n)}^{(a)}$ is defined as

$$\hat{P}_{(n)}^{(a)} = \frac{1}{n_v}\left(\sum_{j=1}^{n_v} d_j\right) \times 100 \tag{3.3a}$$

where n_v is the number of cases in the validation sample. The percentage-N-fold CV (NFCV$_{-P}$) estimate of the actual hit rate, $\hat{p}^{(a)}$ is given as

$$\hat{P}_{NFCV_{-P}}^{(a)} = \frac{1}{N_{-P}} \sum_{n=1}^{N_{-P}} \hat{p}_{(n)}^{(a)} \tag{3.3b}$$

where N$_{-P}$ is the total number of folds based on percentage values of 60, 70, 80, and 90 respectively.

4 The Proposed Variable Selection Method

The new variable selection method which is a modification of the Leave-one-out cross-validation (LOOCV) method [38–40] is proposed to address the problems inherent with the all-possible subset approach. This is aimed at obtaining a subset of predictor variables that is superior both in terms of the number and combination of the predictor variables, as well as the hit rate obtained when used as training sample with less computational expense. This proposed variable selection method helps obtain a superior subset of predictors from the list of already identified potential predictor variables. The outline of the proposed variable selection method is described as follows:

Suppose we are given a data set (or a historical sample, D_N) that consists of N samples $\{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \Re^P$ are the corresponding predictor variable vectors and $y_i \in \{1, 2, \ldots, K\}$ is the group label for the ith sample. Let $D_N = [x_1, x_2, \ldots, x_N] \in \Re^{P \times N}$ be the historical sample data matrix.

1. Using D_N, we build a PDF and obtain its hit rate. The obtained hit rate otherwise known as cutoff hit rate, $P_C^{(a)}$ for this study will serve as an information criterion. If we let

$$d_j = \begin{cases} 1 & \text{if } \hat{Z}_j = Z_j \\ 0 & \text{otherwise} \end{cases}$$

Where $\hat{Z}_j$ is the predicted response for the jth observation in the historical sample, D_N, Z_j is the value of the jth observation in the historical sample, D_N. The cutoff hit rate which serves as an information criterion to determine which variable will be dropped and which variable will be retained is given as

$$P_C^{(a)} = \frac{1}{N}\left(\sum_{j=1}^{N} d_j\right) \times 100 \tag{4.1}$$

where N is the total number of cases over all groups.

2. Next, we removed the first predictor variable, X_1 and obtain its associated hit rate on the remaining variables before replacing it. The obtained hit rate which serves as estimate of variable X_1 individual or unique contribution is denoted by $P_1^{(a)}$. If we let

$$d_j = \begin{cases} 1 & \text{if } \hat{Z}_j^{-X_1} = Z_j \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{Z}_j^{-X_1}$ is the predicted response for the jth observation computed with the first variable removed from the historical sample, D_N, Z_j is the value of the jth observation in the training sample, $D_N^{(t)}$. Variable X_1 associated hit rate, $P_1^{(a)}$ is given as

$$P_1^{(a)} = \frac{1}{N}\left(\sum_{j=1}^{N} d_j\right) \times 100 \tag{4.2}$$

where N is the total number of cases over all groups

3. This process is repeated for the remaining variables in the historical sample, D_N to obtain a list of associated hit rates, $P_1^{(a)}, \ldots, P_P^{(a)}$ corresponding to the unique contributions of the variable list $X_1, \ldots, X_p$ in the historical sample, D_N. Therefore, a total of P + 1 subsets of predictors or PDF_S hit rates need be assessed.
4. If any of the obtained associated hit rates $P_1^{(a)}, \ldots, P_P^{(a)}$ is equal or greater than the cutoff hit rate, $P_C^{(a)}$, the variable associated with that hit rate will be dropped from the analysis, otherwise it is retained. The set of retained variables whose associated hit rates are lower than the cutoff hit rate becomes the best set of predictor variables with significant independent and combined contribution. This

unique set of predictors will now be chosen as the superior subset of predictors (or training sample) from the historical sample, the superior subset is given as

$$D_n = \left[x_1^*, x_2^*, \ldots, x_P^*\right] \tag{4.3}$$

where x_i^* are the variables whose associated hit rates is strictly less than the cutoff hit rate

5 Computational Results

To investigate the performance of the proposed variable selection method, two real world data sets were used. These computational examples consist of comparing our proposed method with that of the all-possible subset approach. All the analysis was done using SPSS 16.

5.1 Case1-a Data Sample Regarding Japanese Banks

Historical sample 1 (HS1) consist of fifty observations for each group. HS1 is from a well-known financial journal among Japanese business leaders which may correspond to Economist, Financial Times, and Business Week in Europe and the United States of America. It involves 100 Japanese financial institutions, along with seven index measures [45]. Each bank is evaluated by the following seven performance indexes:

(1) return on total assets (=total profits/average total assets),
(2) labour profitability (=total profits/total employees),
(3) equity to total assets (=total equity/average total assets),
(4) total net working capital
(5) return on equity (=earnings available for common/average equity),
(6) cost-profit ratio (=total operating expenditures/total profits), and
(7) bad loan ratio (=total bad loans/total loans).

For this data set, we begin the preprocessing step with the all-possible subset method. This approach involves conducting an all possible subsets of each size in order to determine the best subset of predictor variables of any given size, prior to building of the discriminant function in order to maximize hit rate. At the end of the $2^P - 1$ analysis, a subset of five predictor variables was chosen as the best subset. The summary of hit rates for variable subsets from the 2-group Japanese bank data, using the all-possible subset method are presented in Table 1.

In Table 1, the sets of values in column 1 represent the subsets size or number of variables, while the values in column 2 represent the number of all possible subsets for each subset size. The $2^P - 1$ analysis resulted in assessing one hundred and twenty-seven (127) subsets of predictor variables (i.e., the sum of the set of values in column 2). From columns 3 and 4, it appears that the total-group hit rate increases as one, two and three variables (from 78 to 90) are supplemented to the subset of X_3 and X_6. But, as the fourth (X_5) and fifth (X_1) variables are added, the hit rate decreases (from 90 to 87). Going "strictly by numbers", it may be concluded that the subset to be retained is

Table 1 Summary of hit rates for variable subsets

Subset size	Possible subsets	Best subset	Hit rate (%)
2	(21)	$X_3\,X_6$	78
3	(35)	$X_1\,X_6\,X_7$	86
		$X_2\,X_6\,X_7$	86
4	(35)	$X_2\,X_3\,X_6\,X_7$	88
5	(21)	$X_2\,X_3\,X_4\,X_6\,X_7$	90∗
6	(7)	$X_2\,X_3\,X_4\,X_5\,X_6\,X_7$	89
7	(1)	$X_1\,X_2\,X_3\,X_4\,X_5\,X_6\,X_7$	87
Best subset: $X_2\,X_3\,X_4\,X_6\,X_7$			

∗ Best subset with the highest hit rate

Table 2 Summary of predictor variables associated hit rates

Variables	Associated hit rates	Cutoff hit rate (%)
Return on total assets (X_1)	89.0	87
Labour profitability (X_2)	84.0∗	
Equity to total assets (X_3)	85.0∗	
Total net working capital (X_4)	87.0	
Return on equity (X_5)	87.0	
Cost-profit ratio (X_6)	85.0∗	
Bad loan ratio (X_7)	84.0∗	
Best subset: $X_2\,X_3\,X_6\,X_7$		

∗ Variables with significant independent contributions

that excluding X_5 and X_1. Hence, the subset of five variables with asterisked hit rate is the "best subset" to be retained.

A cursory look at Table 1 shows that two subsets of size three yielded the same hit rates. Assuming the subset of size three gave the highest hit rate, the subset of choice may be based on predictor set collection- a researcher judgment call [26].

For the same data set, using our proposed rule as a preprocessing step, the P + 1 analysis resulted in assessing seven (7) subsets of predictor variables. The summary of the seven predictor variables associated hit rates (or individual contributions) and their combined contribution (or cutoff hit rate) are presented in Table 2.

From the results in column 2 of Table 2, four predictor variables with asterisked associated hit rates out of the seven predictor variables were less than the cutoff hit rate, $P_C^{(a)}$ value of 87 %. These four predictor variables that gave the best significant independent and combined contribution becomes the "superior subset" to be retained.

5.2 Case2- a Data Sample Regarding University Demonstration Secondary School (UDSS) Students' Academic Records, University of Benin, Nigeria

Beside the Japanese banks data, we also used a second real data set or historical sample 2 to serve as validation for the first result. Historical sample 2 involves students'

Table 3 Summary of hit rates for variable subsets

Subset size	Possible subsets	Best subset	Hit rate (%)
2	55	$X_2\ X_5$	81.7
3	165	$X_4\ X_5\ X_9$	85.0
		$X_5\ X_8\ X_9$	85.0
		$X_5\ X_9\ X_{10}$	85.0
		$X_5\ X_{10}\ X_{11}$	85.0
4	330	$X_3\ X_5\ X_8\ X_9$	86.7
		$X_4\ X_5\ X_7\ X_{10}$	86.7
5	462	$X_1\ X_3\ X_5\ X_{10}\ X_{11}$	88.3
		$X_4\ X_5\ X_7\ X_8\ X_{10}$	88.3
6	462	$X_1\ X_4\ X_5\ X_7\ X_{10}\ X_{11}$	91.7
7	330	$X_1\ X_2\ X_4\ X_5\ X_7\ X_{10}\ X_{11}$	93.3∗
8	165	$X_1\ X_2\ X_4\ X_5\ X_6\ X_7\ X_{10}\ X_{11}$	93.3
		$X_1\ X_2\ X_4\ X_5\ X_7\ X_9\ X_{10}\ X_{11}$	93.3
9	55	$X_1\ X_2\ X_4\ X_5\ X_7\ X_8\ X_9\ X_{10}\ X_{11}$	91.7
10	11	$X_1\ X_2\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}\ X_{11}$	91.7
11	1	$X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}\ X_{11}$	90.0
Best subset: $X_1\ X_2\ X_4\ X_5\ X_7\ X_{10}\ X_{11}$			

academic records for junior secondary school (JSS) 2, in University Demonstration Secondary School (UDSS), University of Benin, Nigeria [28]. The data set involves average scores for the three terms obtained for eleven (11) subjects which include:

(1) English Language (X_1),
(2) Mathematics (X_2),
(3) Integrated Science (X_3),
(4) Social Studies (X_4)
(5) Introductory Technology (X_5),
(6) Business Studies (X_6)
(7) Home Economics (X_7),
(8) Agricultural Science (X_8),
(9) Fine Art (X_9),
(10) Physical and Health Education (X_{10}), and
(11) Computer Studies (X_{11}).

For this data set, the all possible subset approach gave a best subset of seven variables. The summary of hit rates for all possible variable subsets from the 2-group students' academic records, are presented in Table 3.

In Table 3, the $2^P - 1$ analysis resulted in assessing two thousand and forty-seven (2047) subsets of predictor variables. The result in Table 3 shows that three subsets yielded the same highest hit rates. As a rule of thumb, the subset with the smallest number of predictor variables was chosen. Hence, the subset of seven predictor variables with asterisk hit rate is therefore the best subset.

Table 4 Summary of predictor variables associated hit rates

Variables	Associated hit rates	Cutoff hit rate (%)
English language (X_1)	83.3∗	90.0
Mathematics (X_2)	83.3∗	
Integrated science (X_3)	91.7	
Social studies (X_4)	83.3∗	
Introductory technology (X_5)	81.7∗	
Business studies (X_6)	91.7	
Home economics (X_7)	88.3∗	
Agricultural science (X_8)	90.0	
Fine art (X_9)	90.0	
Physical and Health education (X_{10})	81.7∗	
Computer studies (X_{11})	83.3∗	
Best Subset: $X_1\ X_2\ X_4\ X_5\ X_7\ X_{10}\ X_{11}$		

∗ Variables with significant independent contributions

Using our proposed method as a preprocessing step, at the end of the P + 1 analysis, seven predictor variables associated hit rates were less than the cutoff hit rate, $P_C^{(a)}$ value of 90 %. The summary of the seven predictor variables associated hit rates (or individual contributions) and their combined contribution (or cutoff hit rate) are presented in Table 4.

In Table 4, the P + 1 analysis resulted in assessing twelve (12) subsets of predictor variables. The seven predictor variables with asterisked hit rates becomes the superior subset of predictor variables that gave the best significant independent and combined contribution.

6 Estimation of Actual Hit Rate Based on Historical Sample 1 and 2

The major aim of seeking the subset of best predictors is to maximize hit rate. In PDA assessing the degree of classification accuracy, amounts to estimating a true hit rate [26]. The hit rate obtained by internal classification analysis or simply internal analysis can be expected to be positively biased or spuriously high [26]. The true hit rate estimation process involve seeking answers to two questions: (1) how accurate can a classification rule based on population information be expected to perform? (2) How accurate can a rule based on a given sample be expected to classify individual units in future samples? Consequently, there are essentially two probabilities of correct classification (or two population hit rates). The first is the optimal hit rate, denoted by $P^{(o)}$. This is the hit rate obtained when a classification rule based on known parameters is applied to the population. The second hit rate which is the most widely used is the actual hit rate, denoted by $P^{(a)}$. This is the hit rate obtained by applying a rule based on a particular sample to future samples taken from the same population.

Table 5 Weight estimates and classification accuracy using historical sample 1

Methods function 1	All possible subset method Z_1	Proposed variable selection method Z_2
(Constant)	−6.738	−6.887
Labour profitability (X_2)	0.005	0.002
Equity to total asset (X_3)	0.004	0.005
Total net working capital (X_4)	−0.002	
Cost-profit ratio (X_6)	0.004	0.005
Bad loan ratio (X_7)	0.003	0.003
Prediction accuracy: LOOCV	85.0 %	86.0 %
KFCV	85.0 %	83.0 %
NFCV-$_{-P}$	82.7 %	81.1 %

Unstandardized weights

Table 6 Weight estimates and classification accuracy using historical sample 2

Methods function 1	All possible subset method Z_1	Proposed variable selection method Z_2
(Constant)	−3.681	−3.681
English language (X_1)	−0.043	−0.043
Mathematics (X_2)	0.022	0.022
Social studies (X_4)	−0.043	−0.043
Introductory technology (X_5)	0.130	0.130
Home economics (X_7)	0.038	0.038
Physical and health educ. (X_{10})	−0.084	−0.084
Computer science (X_{11})	0.061	0.061
Prediction accuracy: LOOCV	80.0 %	80.0 %
KFCV	83.0 %	83.0 %
NFCV$_{-P}$	81.1 %	81.1 %

Unstandardized weights

To evaluate the performance of our proposed variable selection method which serves as a preprocessing step prior to building the PDF, an estimate of the actual hit rate was obtained. In order to obtain this estimate, the leave one out CV which is the most classical exhaustive CV procedure, as well as the K-fold CV and percentage-N-fold CV procedures were used. Tables 5 and 6 below shows the SPSS output for discriminant weights obtained for the PDF_S that was built using our method as a preprocessing step, together with that of the all possible subset approach, as well as estimates of the actual hit rate (prediction accuracy) obtained from the leave-one-out cross validation, K-fold CV and percentage-N-fold CV methods

7 Summary and Conclusion

This work approaches the problem of variable selection as a preprocessing step prior to building a PDF in order to maximize hit rate. For historical sample 1 with 7 predictor variables, using the all-possible subset method or the $2^P - 1$ analysis, a total of 127 subsets of predictors (or PDF_S hit rates) were assessed in order to obtain the best subset of predictor variables. For the same historical sample 1 with 7 predictor variables, using our proposed method or the P + 1 analysis, a total of 8 subsets of predictors were assessed in order to obtain the best subset of predictor variables. Also, for historical 2 with 11 predictor variables, a total of $2^P - 1$ (or 2047) subsets of predictors or PDF_S hit rates were assessed using the all-possible subset method. While for the same historical sample 2, a total of P + 1 or 12 subsets of predictors were assessed using our proposed method. The significant reductions in the number of subsets of predictor or PDF_S hit rates that need be assessed suggest that our proposed rule seems to show some reasonable performance in terms of computational expense.

A cursory look at Table 1 shows that two subsets of size three yielded the same highest hit rates. Also, the result in Table 3 shows that three subsets also yielded the same highest hit rates. For the latter, in order to obtain the subset of choice, the subset with minimum number of predictor variable was chosen. Assuming the three subsets that yielded the same hit rate has the same number of predictor variables, additional information criterion would have been needed to obtain the subset of choice, which will amount to additional computational expense. Alternatively, the subset of choice may be based on predictor set collection- a researcher judgment call [26]. However, using the proposed method, the problem associated with having two or more best subsets (i.e., not having a superior subset) of a given size that yields the same hit rate common with all-possible subset method was completely avoided.

In Tables 5, we observed that the LOOCV estimate of PDF's hit rate obtained using our proposed rule was slightly higher than that of the all possible subset method. In Table 6, we also observed that both methods had the same predictive performance in terms of their actual hit rates. Among the three CV methods that was used in obtaining estimates of the actual hit rate, the LOOCV method is known to be approximately unbiased, and is often used in assessing the performance of classifiers [46,47]. Since our rule was able to achieve this feat with significantly less computation expense, truly shows that our propose rule seems to perform better. Therefore, using our proposed rule, we need just P + 1 subset of predictors that must be considered irrespective of the number of predictor variables.

Although the results support the applicability and the merit of the proposed rule, however, it is true that this verification is limited to the scope of the data sets used. Therefore, this article believes that more experimental results are still called for in order to make a final conclusion on the superiority of the proposed rule over a variety of known classical alternatives. That is another future research task. Finally, it is hoped that this research makes a small contribution to research works on variable selection as a preprocessing step in PDA.

Acknowledgments The authors would like to thank the reviewers for their valuable comments which contributed to an improved version of this paper.

References

1. Huberty CJ (1994) Applied discriminant analysis. Willey, New York
2. Thomas BM, Nema D, Adrian ER (2010) Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. Ann Appl Stat 8(1):369–421
3. Reunanen J (2003) Overfitting in making comparisons between variable selection methods. J Mach Learn Res 3(7/8):1371–1392
4. Lewis PM (1962) The characteristics selection problem in recognition systems. IEEE Trans Inf Theory 8:171–178
5. Ganster H, Pinz A, Rohrer R, Wildling G, Binder M, Kittler H (2001) Automated melanoma recognition. IEEE Trans Med Imaging 20(3):233–239
6. Inza I, Sierra B, Blanco R (2002) Gene selection by sequential search wrapper approaches in microarray cancer class prediction. J Intell Fuzzy Syst 12(1):25–33
7. Shy S, Suganthan PN (2003) Feature analysis and classification of protein secondary structure data. Lect Notes Comput Sci 2714:1151–1158
8. Tamoto E, Tada M, Murakawa K, Takada M, Shindo G, Teramoto K, Matsunaga A, Komuro K, Kanai M, Kawakami A, Fujiwara Y, Kobayashi N, Shirata K, Nishimura N, Okushiba SI, Kondo S, Hamada J, Yoshiki T, Moriuchi T, Katoh H (2004) Gene expression profile changes correlated with tumor progression and lymph node metastasis in esophageal cancer. Clin Cancer Res 10(11):3629–3638
9. Chiang LH, Pell RJ (2004) Genetic algorithms combined with discriminant for key variables identification. J Process Control 14:143–155
10. Pacheco J, Casado S, Nunez L, Gomez O (2006) Analysis of new variable selection methods for discriminant analysis. Comput Stat Data Anal 51:1463–1478
11. Louw W, Steep SJ (2006) Variable selection in kernel fisher discriminant analysis by means of recursive feature elimination. Comput Stat Data Anal 51:2043–2055
12. Trendafilov NT, Jolliffe IT (2007) DALASS: valriable selection in discriminant analysis via the LASSO. Comput Stat Data Anal 51:3718–3736
13. Osemwenkhae JE, Iduseri A (2011) Efficient data-driven rule for obtaining an optimal predictive function of a discriminant analysis. J Niger Assoc Math Phys 18:373–380
14. Mary-Huard T, Robin S, Daudin JJ (2007) A penalized criterion for variable selection in classification. J Multivar Anal 98:695–705
15. Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc, Ser B 58:267–288
16. Bertrand C, Ernest F, Hao HZ (2009) Principles and theory for data mining and machine learning. Springer series in statistics doi:10.1007/978-0-387-98135-2, Springer Science+Business Media New York, pp 569–576
17. Chiang LH, Russell EL, Braatz RD (2001) Fault detection and diagnosis in industrial systems. Springer, New York
18. Kong S, Wang D (2012) A brief summary of dictionary learning based approach for classification. arxiv.org/pdf/1205.6544
19. Yang M, Zhang J, Yang J, Zhang D (2010) Metaface learning for sparse representation based face recognition. In: Proceedings of the IEEE international conference on image processing (ICIP), pp 1601–1604
20. Ramirez I, Sprechmann P, Sapiro G (2010) Classification and clustering via dictionary learning with structured incoherence and shared features. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 3501–3508
21. Mairal J, Bach F, Ponce J, Sapiro G, Zisserman A (2008) Supervised dictionary learning. Advances in neural information processing systems 21 (NIPS). See http://papers.nips.cc/paper/3448-supervised-dictionary-learning.pdf
22. Zhang Q, Li B (2010) Discriminative K-SVD for dictionary learning in face recognition. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 2691–2698
23. Jiang Z, Lin Z, Davis LS (2011) Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 1697–1704
24. Yang M, Zhang L, Feng X, Zhang D (2011) Fisher discrimination dictionary learning for sparse representation. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 543–550

25. Welge P (1990) Three reasons why stepwise regression methods should not be used by researchers. Paper presented at the annual meeting of the southwest educational research association, Austin, TX, January 1990
26. Huberty CJ, Olejnik S (2006) Applied manova and discriminant analysis. Wiley, Hoboken
27. Gareth J, Daniela W, Trevor H, Robert T (2013) An introduction to statistical learning: with application in R. Springer Texts in Statistics 103: doi:10.1007/978-1-4614-7138-7, Springer Science+Business Media, New York, pp 205–207
28. Iduseri A (2015) A two-step training sample optimization rule for the application of two and multiple group discriminant analysis. Dissertation, University of Benin, Benin City, Nigeria
29. Efroymson MA (1960) Multiple regression analysis. In: Ralston A, Wilf HS (eds) Mathematical methods for digital computers. Wiley, New York, pp 191–203
30. Draper NR, Smith H (1981) Applied regression analysis. Wiley, New York
31. Thompson B (1995) Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. Educ Psychol Meas 55(4):525–534
32. Klecka WR (1980) Discriminant analysis. Quantitative application in social sciences series, vol 19. Sage Publications, Thousand Oaks
33. Whitaker JS (1997) Use of stepwise methodology in discriminant analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin
34. Mosteller F, Tukey JW (1968) Data analysis including statistics. In: Lindzey G, Aronson E (eds) Handbook of social psychology, 2nd edn. Addison-Wesley, Reading, pp 80–203
35. Morrison DG (1969) On the interpretation in discriminant analysis. J Mark Res 6(May):156–163
36. Devroye L, Wagner TJ (1979) Distribution-free performance bounds for potential function rules. IEEE Trans Inf Theory 25(5):601–604
37. Bartlett PL, Boucheron S, Lugosi G (2002) Model selection and error estimation. Mach Learn 48:85–113
38. Stone M (1974) Cross-validation choice and assessment of statistical predictions. J R Stat Soc Ser B 36:111–147 With discussion and a reply by the authors
39. Allen DM (1974) The relationship between variable selection and data augumentation and a method for prediction. Technometrics 26:125–127
40. Geisser S (1975) The predictive sample reuse method with applications. J Am Stat Assoc 70:320–328
41. Iduseri A, Osemwenkha JE (2014) On estimation of actual hit rate in the categorical criterion predicting process. J Niger Assoc Math Phys 28(1):461–468
42. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth statistics/probability series. Wadsworth Advanced Books and Software, Belmont
43. Breiman L, Spector P (1992) Submodel selection and evaluation in regression. The x-random case. Int Stat Rev 60(3):291–319
44. Iduseri A, Osemwenkhae JE (2014) On estimation of actual hit rate in the categorical criterion predicting process. J Niger Assoc Math Phys 28(1):461–468
45. Sueyoshi T (1999) DEA-discriminant analysis in the view of goal programming. Eur J Opl Res 115:564–582
46. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: 14th international joint conference on artificial intelligence (IJCAI), Montreal
47. Isler Y, Narin A, Ozer M (2015) Comparison of the effects of cross-validation methods on determining performances of classifiers used in diagnosing congestive heart failure. Meas Sci Rev 15(4):196–201