Journal of the Nigerian Association of Mathematical Physics
Volume28, No. 1, (*November, 2014*), pp 461 – 468

On Estimation of Actual Hit Rate in the Categorical Criterion Predicting Process

Iduseri A. and Osemwenkhae J. E.

Department of Mathematics, University of Benin, Nigeria.

Abstract

This paper examines a new approach of using the percentage-N-fold cross validation rule ($NFCV_{-P}$) to determine the expected value for the actual hit rate of a predictive discriminant function (PDF). The method is similar to K-fold cross validation method (KFCV) used to predict the fit of a model to a hypothetical validation set when an explicit validation set is not available. The new approach produces estimates of the actual hit rate that were consistent, and essentially equivalent to that of the leave-one-out cross validation (LOOCV) method with less computational expense.

Keywords: Percentage-N-fold cross-validation, Actual hit-rate, Predictive discriminant function, Categorical criterion.

1.0 Introduction

When the criterion for prediction involves one or more predictor variables along side with a categorical criterion, such prediction will call for the use of predictive discriminant analysis (PDA). In PDA, assessing the degree of accuracy of a predictive discriminant function (PDF) amounts to estimating a true hit rate (in particular the actual hit rate, denoted here by $P^{(a)}$). This is the hit rate obtained by applying a rule based on a particular training sample to future samples taken from the same population. That is, $P^{(a)}$ may be thought of as the expected proportion of correct classifications over future samples yielded by a rule based on statistics from a particular training sample.

Among the several notable cross-validation (CV) methods used in estimation of $P^{(a)}$, leave-one-out cross-validation (LOOCV) is the most frequently used one, because its estimation is considered a nearly unbiased estimate of $P^{(a)}$. It is the most classical exhaustive CV procedure originally designed to estimate $P^{(a)}$. However in the case of large training sets, LOOCV method can be computationally intractable. Also, due to the reuse of the original data, LOOCV method often yields relatively high varied results [1, 2]. The increased variance may be due to the fact that the intersection between the complements of two holdout portions has n-2 data points. These data points are used, along with the one extra point, in fitting the PDF to predict the point left out. Thus, the PDF is fit twice on almost the same data, giving highly dependent predictions; dependence typically inflates variance [1, 3, 4, 5]. K-fold cross-validation (KFCV) was then introduced in [6] as an alternative to the computationally expensive LOOCV method. The work in [7] found 5-fold and 10-fold CV to work better than LOOCV method, while the work in [3] suggested k should be chosen between 5 and 15, depending on n, how to determine a good choice for "k" has remained a major drawback, and a handy rule of thumb is yet to be established. Generally, in literature a value of 10 for k is popular for estimating the error rate of $P^{(a)}$.

The task of estimating the $P^{(a)}$ (i.e. the predictive accuracy) of the PDF for the data on hand, thereby obtaining the fit of the PDF to a hypothetical validation set (i.e., two or more validation samples), has received much theoretical attention. Over the years, statisticians and methodologists have resorted to hit rate estimator methods such as *formula method* [8, 9, 10], *internal analysis method* [11, 12] and the *external analysis methods* which include validation procedure and Cross-validation procedure to accomplish this task. Notable variants of the validation procedure include hold-out sample method [13, 14] and the repeated random sub-validation method [15]. Other notable variants of the cross-validation procedure include leave-one-out cross- validation method [13, 16], and the K-fold cross-validation method [6, 7]. The use of formula method for estimating the $P^{(a)}$, has been restricted to the two-group multivariate normal case [10]. The formula method is generally recognized as a poor estimate of $P^{(a)}$ [1]. The observed hit rates obtained using the internal analysis method may be misleading, since in most cases the observed hit rate is spuriously high [12]. Similarly, the external analysis methods, which include validation procedure and cross-validation procedure generally, yield good estimators in the sense of accuracy, yet they are quite uneconomical with real data set [10, 12]. The results of two simulation studies [4, 5], indicates that the external analysis methods may yield high hit rate estimates that have relatively

Corresponding author: ***Iduseri A.,*** E-mail: augustine.iduseri@uniben.edu, Tel.: +2348036698860

high variability over repeated sampling. This relatively high variability may be due to the reuse of the original data.
In this paper, a modified CV procedure is developed by adapting the K-fold cross-validation (KFCV) method introduced in [6], as an alternative to the computationally expensive LOOCV method. This variant of CV builds a CV procedure by changing the choice of "k"(i.e., the number folds) to a known decision variable. This paper will show how the modified CV procedure otherwise known as the percentage-N-fold cross-validation rule, $NFCV_P$ [17] can be used to obtain a true hit rate estimate of a PDF that is essentially equivalent to that of the LOOCV method with less computation time.

2.0 Estimation of Actual Hit Rate

To describe the propose rule otherwise known as the percentage-N-fold cross-validation rule ($NFCV_{-P}$), this paper first returns to LOOCV method pioneered by Stone [13] and the KFCV method originally introduced by Geisser [6] which are notable variants of the *external analysis method*.

2.1 *Leave-One-Out Cross-Validation (LOOCV) Method*

The LOOCV method [13] was originally designed to estimate the actual hit rate, $P^{(a)}$. It involves using a single observation from the *historical* sample, denoted here as D_N (this is the complete list of objects over all groups) as the validation data, and the remaining observations as the training set. This process is repeated N times such that each observation in D_N is used once as the validation data, and the proportions of deleted units (test samples) correctly classified are used as the hit-rate. If we let

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

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 $P^{(a)}$ is given by

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

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

2.2 *K-Fold Cross-Validation (KFCV) Method*

K-fold CV was introduced by Geisser [6] as an alternative to the computationally expensive LOOCV method [18]. In K-fold cross-validation method (KFCV), the historical sample, D_N is partitioned into K sub-samples (folds). Of the K sub-samples, K-1 fold is used for training and the remaining one for testing. The cross-validation process is then repeated K times, with each of the k sub-samples used for validation exactly once. If we denote $P^{(a)}$ as the hit rate for each of the k sub-samples, and let

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

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 D_N, $k(j)$ is the fold containing observation j. Then the hit rate for the k sub-sample is define as

$$P_k^{(a)} = \frac{1}{n_V}\sum_{j=1}^{n_V} d_j \times 100 \tag{2}$$

and n_v is the number of cases in the validation sample. Therefore, the KFCV estimate of $P^{(a)}$ is given as

$$\hat{P}_{KFCV}^{(a)} = 1/K \sum_{k=1}^{K} P_k^{(a)} \tag{3}$$

2.3 The Proposed Rule

The proposed rule we present in this work, which is a modification of the K-fold cross-validation method involves carrying out a percentage-N-fold partitioning of the data set and will therefore be referred to as the percentage-N-fold partitioning rule (NFCV$_{-P}$). Let $D_N = [x_1, x_2, \ldots, x_P]$ be the historical data matrix and $D_k \in \Re^{n_k}$ be the data matrix of the kth group, where n_k is the sample size of the kth group, and $\sum_{k=1}^{K} n_k = N$. From the historical sample, D_N, we compute the PDF, Z given as

$$Z = u_1 X_1^* + u_2 X_2^* + \ldots + u_p X_p^* = \eta\left(D_n^t\right) \tag{4}$$

where $Z_{(opt)}$ is the linear discriminant function, u_i are the discriminant weights, X_i^* are the selected predictor variables and $\eta\left(D_n^*\right)$ indicates that the PDF is calibrated on selected predictor variable from a pool of identified potential predictor variables.

In order to obtain the a true hit rate estimate of PDF, Z, we begin by twenty (20) validation sets, $I^{(v)}$ from any given historical sample using the outline of NFCV$_{-P}$ rule 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 validation sample, $D_n^{(v)}$

Step3: Repeat steps 1-2 using percentage values of 60, 70, 80 and 90 respectively (see Appendix A for percentage–N-fold partitioning of data set).

If we denote $\hat{P}_{(n)}^{(a)}$ as the estimate of $P^{(a)}$ for each of the n Validation samples, and let

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

where $\hat{Z}_j^{-n(j)}$ is the predicted response for the jth observation computed with the n(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

$$\hat{P}_{(n)}^{(a)} = \frac{1}{n_V}\sum_{j=1}^{n_V} d_j \times 100 \tag{5}$$

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

$$\hat{P}_{NFCV_{-P}}^{(a)} = \frac{1}{N_{-P}}\sum_{n=1}^{N_{-P}} \hat{p}_{(n)}^{(a)} \tag{6}$$

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

3.0 Computational Examination and Results

3.1 Historical Samples 1 and 2

Two historical samples (or data sets) were used for this study: Historical Sample 1 [19] includes overall grade point average (GPA) for 100 level and grades in all statistics and mathematics core courses from students' academic records for 100 and 200 levels, in the Department of Statistics, in a University system, while Historical Sample 2 [17] includes measures of interest in outdoor activity, sociability and conservativeness of employees in three different job classifications (customer service personal, mechanics, and dispatchers) from a large international air carrier. Since PDA is concerned with hit rates and accuracy of classification, and any reasonable PDA stepwise procedures must focus on

maximizing hit rates. To confirm the GPA and STA 201 (for Historical 1) as well as outdoor activity, sociability and conservativeness (for historical 2) as the best subsets of the predictor variables using the forward stepwise analysis, we then use an "all-possible-subsets" approach [20, 21] which gave the same result with that of the forward stepwise analysis.

3.1.1 Methodological alternatives

Using the two historical samples as shown in Appendix 1 and 2, this study examines the performance of three different $P^{(a)}$ estimation methods in PDA, including

(a) LOOCV (leave-one-out cross-validation) [6, 13, 22]
(b) KFCV (V-fold cross-validation) [6, 20]
(c) NFCV$_{-P}$ (Percentage-N-fold cross-validation) proposed in this paper.

The first and second hit-rate, $P^{(a)}$ estimation methods are well known CV procedures and the first one is incorporated in many computer software. Also, the first two methods (LOOCV and KFCV) produce more than one PDF. For the LOOCV method, the number of PDF$_S$ is equal to the size of the training sample used, while for the KFCV method, the number of PDF$_S$ is equal to the number of folds used. (Hence, this study omits results for the obtained PDF$_S$ for both methods as well as the summary of hit-rates for the LOOCV method, hereafter.)

For the proposed rule, using the outline of the percentage-N-fold cross-validation procedure, we obtain for Historical sample 1 and 2 the PDF$_S$ given as

$$Z_1 = u^*_1 X_1 + u^*_2 X_2$$
$$= 1.277\,(GPA) + 0.045\,(STA201) \quad (7)$$
$$Z_2 = u^*_1 X_1 + u^*_2 X_2 + u^*_3 X_3$$
$$= 0.209\,(OUTDOOR\ ) - 0.100\,(SOCIAL\ ) + 0.120(CONSERV\ ) \quad (8)$$

3.1.2 Computational results

Table 1: Summary of Hit Rate Results

Historical sample 1			Historical sample 2		
KFCV hit rates	NFCV$_{-P}$ hit rates		KFCV hit rates	NFCV$_{-P}$ hit rates	
5 folds	10 folds	20 folds	5 folds	10 folds	20 folds
95.8	66.7	87.5	78.6	71.4	82.1
91.7	100.0	87.5	85.7	78.6	82.1
91.7	75.0	83.3	85.7	78.6	81.0
66.7	100.0	97.2	78.6	85.7	85.7
87.5	100.0	80.6	82.1	92.9	81.0
	83.3	83.3		78.6	78.6
	66.7	91.7		85.7	85.7
	83.3	95.8		71.4	85.7
	83.3	79.2		85.7	78.6
	91.7	87.5		78.6	82.1
		66.7			71.4
		100.0			85.7
		83.3			85.7
		100.0			85.7
		100.0			92.9
		91.7			78.6
		66.7			64.3
		91.7			71.4
		83.3			85.7
		91.7			78.6

Table 1 documents hit rate results for two hit rate estimation methods (K-fold cross-validation method and our proposed method otherwise known as the percentage-N-fold cross-validation rule) obtained from validation sets from historical sample 1 and 2. The first two columns under historical sample 1 list the hit rate results for the 5-fold and 10-fold variants of KFCV method; while the third column lists the hit rate results for the NFCV$_{-P}$ rule. Similarly, the first two columns under historical sample2 list the hit rate results for the notable 5-fold and 10-fold variants of KFCV method, while the third column lists the hit rate results for the NFCV$_{-P}$ rule.

Table 2: Summary of Actual Hit Rate Estimates for LOOCV, KFCV and NFCV$_{-P}$ Methods

	LOOCV	KFCV		NFCV$_{-P}$	ValidN (listwise)
		5 fold	10 fold		Unweighted
Historical sample 1 **(87.5)**	86.7	86.7	85.0	87.4	120
Historical sample 2 **(81.4)**	80.7	82.1	80.7	81.1	140

In Table 2, we present the summaries of the actual hit rate, $P^{(a)}$ for the three hit rate estimation methods considered in this study. The estimates of the actual hit rate, $P^{(a)}$ shown in Table 2, are simply the average values of the various hit rate results listed in the six columns of Table 1 (except that of LOOCV method that was obtained directly from SPSS 16 statistical package output result) . Also, the two hit rates results shown in column 1 of Table 2 whose values are written in bold cases represent the hit rates obtained when both historical samples were used as validation samples.

4.0 Discussion of Results

To obtain the two LOOCV actual hit rate, $P^{(a)}$ estimate in column 2 of Table 2, a total of 120 and 140 validation samples as well as 120 and 140 PDF$_S$ respectively are needed. Similarly, for the 5-fold and 10-fold variants of KFCV method, a total of 5 and 10 validation samples as well as 5 and 10 PDF$_S$ respectively are needed to obtain the actual hit rate, $P^{(a)}$ estimate in column 3 of Table 2. While the NFCV$_{-P}$ actual hit rate, $P^{(a)}$ as shown in column 4 of Table 2, requires 20 validation samples and 20 PDF$_S$ respectively.

In column 3 of Table 2, under the 5-fold variant of the KFCV method, the actual hit rate, $P^{(a)}$ estimate of 86.7 for historical sample 1 is equal to that of the LOOCV estimate in column 2 of Table 2, while the 5-fold variant estimate of 82.1 for historical sample 2 is significantly different from that of the LOOCV estimate. But still in column 3 of Table 2, under the 10-fold variant, the reverse was the case. This finding indicates that the 5-fold variant of the KFCV performs better than the 10-fold variant for historical sample 1 since its estimate was equal to that of the LOOCV estimate, while the 10-fold variant performs better than the 5-fold variant for historical 2. Consequently, one need to have an idea of the LOOCV hit rate estimate, $P^{(a)}$ to serve as a guard in determining a good choice of "K" (number of folds) when using the KFCV method. In addition, this finding also indicates that the KFCV method actually has a drawback in determining the choice of "K" or the variant to use. Breiman and Spector [7] have already reported that the choice of "K" has remained a major drawback.

A cursory look at our proposed rule (NFCV$_{-P}$) estimate of $P^{(a)}$ in Table 2 shows that our $\hat{P}^{(a)}$ estimate of 87.4 for historical sample 1 and 81.1 for historical sample 2 are essentially equivalent to that of the LOOCV estimates of 86.7 and 80.7 respectively. Also, when both historical sample 1 and historical sample 2 were used as validation samples (otherwise known as internal analysis method), the hit rate estimates of 87.5 for historical 1 and 81.4 for historical sample 2 are also essentially equivalent to our proposed rule actual hit rates estimates. Since we are able to obtain actual hit rate estimates that are not only consistent when compared with that of KFCV method, but essentially equivalent to LOOCV estimate with less computation expense, one can maintain that our method proves better

Also this study has also reveal the need to have an idea of the LOOCV hit rate estimate, $P^{(a)}$ to serve as a guard in determining a good choice of "K" (number of folds) when using the KFCV method.

5.0 Conclusion

Thisstudy has examined categorical criterion predicting process (in particular, PDA), especially in the area of assessingthe predictive accuracy of a PDF which amounts to estimating the actual hit rate, $\hat{P}^{(a)}$. The new approach produces estimates of the actual hit rates that were consistent, and essentially equivalent to that of the LOOCV method with less computational expense.

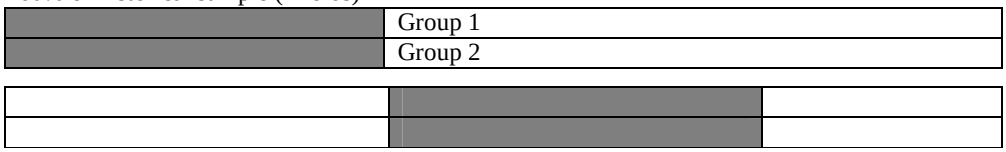
The efficiency of the proposed rule is determined by comparing it with classical CV variants using two real data sets. Another important result of this study is in providing an alternative choice of k (i.e., 20 folds), which yielded a consistent and better estimate of the actual hit rate compared to that of the K-fold CV variants. It must be stated that the calculation time of our method is greater: in this work 20 folds was needed to obtain an estimate for the actual hit rate, versus the 5 or 10 folds used in K-fold CV variants. However, it is clear that for most applications and studies this has no special relevance if we take into account the advantages presented in the foregoing.

Finally, two real historical samples have been used. Consequently, the validity of the experimental results is limited to the scope of the data sets used. Therefore, this article believes that more experimental results are called for in order to make a final conclusion on the efficiency of the proposed rule over the known classical alternatives.

Appendix A

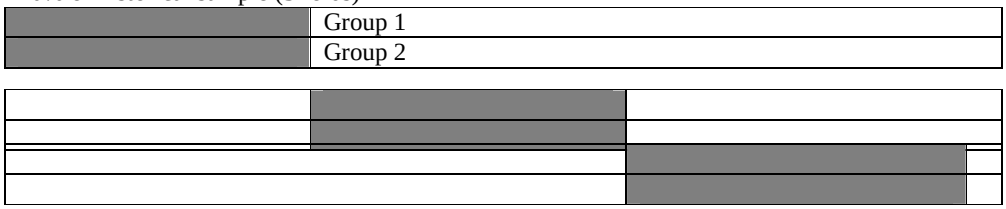
Percentage-N-fold Partition of Historical Sample of Size 120 for Two Group

60% of historical sample (2 folds)



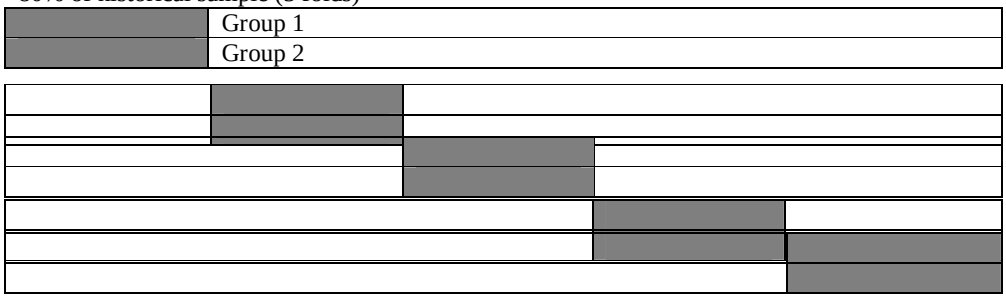
1. Shaded parts = deleted data points of size 48
2. Unshaded parts = training samples of size 72

70% of historical sample (3 folds)



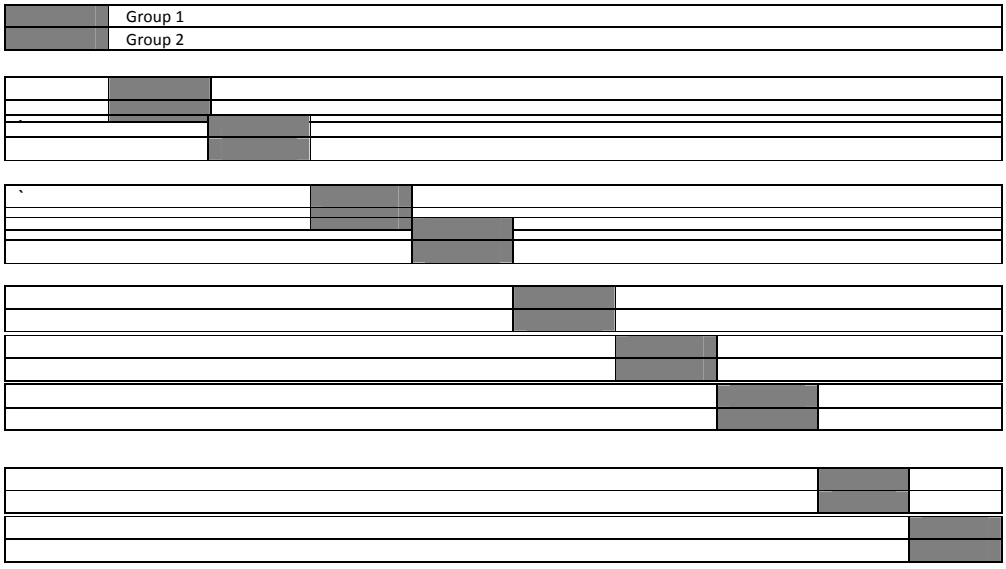
1. Shaded parts = deleted data points of size 36
2. Unshaded parts = training samples of size 84

80% of historical sample (5 folds)



1. Shaded parts = deleted data points of size 24
2. Unshaded parts = training samples of size 96

90% of historical sample (10 folds)



1. Shaded parts = deleted data points of size 12
2. Unshaded parts = training samples of size 108.

References

[1] Huberty, C. J., and Olejnik, S. (2006). Applied Manova and Discriminant Analysis.Hoboken, New Jersey. John Wiley and Sons, Inc.

[2] Gareth, J., Daniela, W., Trevor, H., and Robert T. (2013). An introduction to statistical learning: with application in R, Springer Texts in Statistics 103, D0I 10.1007/978-1-4614-7138-7, Springer Science+Business Media New York, 205- 207.

[3] Bertrand, C., Ernest, F., Hao Helen, Z. (2009). Principles and Theory for Data mining and Machine Learning. Springer Science+Business Media, LLC, 233 Springer Street, NewYork, NY 10013, USA.

[4] Glick, N. (1978). Additive Estimators for Probabilities of Correct classification. Pattern Recognition, 10(297, 302, 303): 211-227.

[5] Hora, S. C., and Wilcox, J. B. (1982). Estimation of error rates in several population discriminant analysis, Journal of Marketing Research, 19(302,303,304,336): 57-61

[6] Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. J.Amer. Statist. Assoc., 70: 320-328

[7] Breiman, L. and Spector, P. (1992). Submodel Selection and Evaluation in Regression. The x-random case. International Statistical Review, 60(3): 291-319

[8] Huberty, C. J., and Mourad, S. A. (1980). Estimation in Multiple Correlation/Prediction Educational and Psychological Measurement, 40 (299, 302):101-112

[9] Dorans, N. J. (1988). The Shrinken Generalized Distance Estimator of the Actual Error Rate in Discriminant Analysis. Journal of Educational Statistics, 3 (298, 358): 63-74

[10] Mclachlam, G. J. (1992). Discriminant Analysis and Statistical Pattern Recognition, New York: Wiley.

[11] Mclachlam, G.J. (1977). Estimating the Linear Discriminant Function from Initial Samples Containing a Small Number of Unclassified Observations. Journal of the Americans Statistical Association 72: 403-406.

[12] Hand, D. J. (1997). Construction and Assessment of Classification Rules: Chicester, UK: Wiley. (12,300,302,311,336,346,366,369,381,386,395,403,409)

[13] Stone, M. (1974). Cross-Validation Choice and Assessment of Statistical Predictions. J. Roy. Statist. Soc. Ser. B, 36. Pg. 111-147. With discussion and a reply by the authors. MRO35637

[14] Devroye, L. and Wagner,T. J. (1979). Distribution-Free Performance Bounds for Potential Function Rules. IEEE Transaction in Information Theory, 25(5): 601-604. MR0545015

[15] Wold, S. (1978). Cross-validation Estimation of the Number of Components in Factor and Principal Component Model., Technometrics, 20: 397-405 [ISI]

[16] Li, K.C. (1987). Asymptotic Optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann. Statist., 15(3): 958-975. MR0902239

[17] Iduseri, A., and Osemwenkhae, J. E. (2011). On Estimation of Discriminant Weights with Best Matching Performance. Journal of Engineering Science and Applications (JESA), 7(2): 32-42

[18] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

[19] Erimafa, J. T., Iduseri, A. and Edokpa, I. W. (2009). Application of Discriminant Analysis to Predict the Class of Degree for Graduating Students in a University System. International Journal of Physical Sciences. Vol. 4(1), pp. 016-021.

[20] Huberty, C.J. (1989). Problems with Stepwise Methods: Better Alternatives. In B.Thompson (ED), Advances in Social Science Methodology,1:43-70.Greenwich, CT: JIA Press.

[21] Thompson, B. (1995). Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply Here: A guidelines editorial. Educational and Psychological Measurement, 55(4): 525- 534.

[22] Allen, D. M. (1974). The Relationship Between Variable Selection and Data Augumentation and a Method for Prediction. Technometrics, 16: 125-127. MRO343481