British Journal of Mathematics & Computer Science
5(5): 606-612, 2015, Article no.BJMCS.2015.045
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Supervised Linear Classification Performance Based on Marginal Probability for Two Groups

F. Z. Okwonu[1*]

[1]Department of Mathematics and Computer Science, Delta State University, Abraka, Nigeria.

Article Information

DOI: 10.9734/BJMCS/2014/13449
Editor(s):
(1) Yilun Shang, Department of Computer Science and Institute for Cyber Security, University of Texas at San Antonio, USA.
Reviewers:
(1) Anonymous, Southeast University, China.
(2) Paweł Karczmarek, Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin, Poland.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=729&id=6&aid=7053

Short Research Article

Received: 18 August 2014
Accepted: 05 September 2014
Published: 21 November 2014

Abstract

The conventional technique to determine classification performance for the linear classification techniques strictly depends on the mean probabilities of correct classification or misclassification. Based on the mean probabilities of correct classification, robustness can be determined. In this paper, a new analytic procedure based on the joint and marginal probabilities is applied to determine robustness and the number of sample observations correctly classified. The classification results computed using this approach is unbiased. This technique is applied to investigate the classification performance of the Fisher linear classification analysis and the robust Fisher's technique based on the minimum covariance determinant. The performance analysis when compared to the conventional procedure revealed that this technique is very informative. Relying on the analysis and the data set used, the recognition rate of the conventional approach is more accurate than the robust Fisher's technique.

Keywords: Classification, mean probability, joint probability, marginal probability.

1 Introduction

The Fisher linear classification analysis (FLCA) [1], is a dimension reduction technique used prior to classification and discrimination [2]. Conventionally, the FLCA procedure was proposed for

**Corresponding author: fzokwonu_delsu@yahoo.com;*

two groups. Its basic advantage is the graphical representation in two dimensions. The FLCA assist in gaining information regarding the separation between the two groups with regards to the within group mean and the contribution of the profile variables [3,4]. It belongs to the class of supervised linear classification technique [5-6]. The basic assumption of the FLCA is the equality of the within group covariance matrices. The coefficient of the FLCA is computed based on the difference between the within group mean vectors and the pooled common covariance matrix. The sample mean and covariance estimates are the building blocks of the FLCA but are sensitive to influential observations [7-12-15]. The sample mean vectors and covariance matrices computed based on data set generated from a multivariate normal distribution enhances the performance of the FLCA maximally [16-17]. On the other hand, if the data set is not drawn from a multivariate normal distribution, the sample mean and covariance estimates computed are influenced by influential observations hence when these sample estimates are applied to develop the FLCA; the misclassification rate for the FLCA tends to increase maximally.

It has been suggested that when the data set are not normally distributed the sample estimates are influenced by influential observations, hence various propositions have been proposed to robustify the sample estimates to enhance maximum classification performance. The maximum likelihood estimator (*M* estimator) [18], generalized maximum likelihood estimator (*GM* estimators) [19], Smooth estimator (*S* estimator) [20], minimum volume ellipsoid (*MVE*) [21] and the minimum covariance determinant estimator (*MCD*) [22] were proposed to robustify the sample mean and covariance matrices. The robustified mean vectors and covariance matrices are substituted into the conventional Fisher linear classification technique to obtain robust Fisher's classification technique. The MCD procedure has been applied to robustify the Fisher linear discriminant analysis and the quadratic discriminant analysis [23]. The MCD procedure strictly depends on information glean from the half set. Detail of this robust high breakdown method and its application to classification is contained in [24]. This paper is concerned on methods to determine robustness and the number of sample sizes correctly classified or misclassified. The conventional approach applies the mean probability of correct classification or the apparent error rate using information from the confusion matrix. In this paper we apply a new technique to determine robustness and the number of sample observations correctly classified using joint and marginal probabilities respectively.

The remainder of this paper is organized as follows. The Fisher linear classification analysis is described in Section Two. Section Three contains robust Fisher linear classification analysis based on minimum covariance determinant. Section Four describes the performance of linear classification techniques. Simulation and conclusions are described in Sections Five and Six, respectively.

2 Conventional Fisher Linear Classification Analysis (FLCA)

The Fisher linear classification analysis [1] for two groups problem is defined mathematically as follows,

$$h\_pv = u^T x, \tag{1}$$

where u denotes the Fisher linear coefficient, x is the sample observation and h_pv denotes the Fisher's classification sore, a scalar. The Fisher's technique is a linear combination of the

observed variables that best describes the maximum separation between the groups [5]. Since the population mean vectors and covariance matrices are unknown, the sample estimates are used to estimate the population mean vectors and covariance matrices respectively. The estimate of the population covariance matrix is unbiased and the evaluation of the Fisher's linear classification scores based on the group mean vectors and the difference between the mean of the Fisher's linear classification score is approximately the Mahalonobis distance [6].

The following equation in comparison with the classification score allows an observation to be assigned to the correct group, say,

$$mean\_cut = \frac{\sum_{i=1}^{2} \overline{x}_i}{2} u^T. \quad (2)$$

Where $mean\_cut$ denote the midpoint and $\overline{x}_i$ are the within group mean vectors. The computation of the Fisher linear coefficient is possible if the group means are unequal. To design the allocation rule for the two groups based on the multivariate sample observations, let $\beta_i (i = 1, 2)$ denote the prior probabilities for the two groups and let us assume that $\beta_1 = \beta_2$ with the basic understanding that $\sum_{i=1}^{2} \beta_i = 1.$ Define $\eta_{c1} = \varpi(2/1)$ to be the cost of misallocating an observation in group two into group one and let $\eta_{c2} = \varpi(1/2)$ be the cost of misallocating an observation from group one into group two, respectively. The total probability of misallocation is given as $\Omega = \beta_1 \eta_{c1} + \beta_2 \eta_{c2}.$ The total probability of correct allocation is obtained by taking the sum of the diagonal of the confusion matrix divided by the total sample size and the misallocation probability otherwise. In practice, the cost of misallocation is not known; hence Fisher's allocation rule is based on the assumption that the prior probabilities and misallocation cost for both groups are equal. The comparison between the classification score and the midpoint defines the linear classification rule. The Fisher linear classification rule is obtained by comparing the classification score with the classification midpoint. The allocation rule is based on Equations (1-2). An observation is assigned to group one if the classification score is greater than or equal to the midpoint otherwise the observation is assigned to group two if the classification score is less than the midpoint. Interestingly, the technique discussed above underperforms if the data set contains influential observations. In order to enhance the performance of the FLCA, robust procedures based on clean data was proposed. The following section three gives detailed account of robust technique based on the minimum covariance determinant estimators.

3 Robust Fisher Linear Classification Analysis Based on Minimum Covariance Determinant (FMCD)

In general, robust linear classification procedures are based on weighting approach, say assigning zero to influential observations and one to regular observations. Specifically, there are no basic rules on applying the weighting technique on the training or test data set or weighting before

splitting the data set into training/testing. Relying on the weighting technique a less technical robust procedure computes its estimates (mean vectors and covariance matrices) and applies these estimates to the conventional FLCA. In this consideration, a more technical approach based on the minimum covariance determinant is considered. This procedure is described based on the half set. The minimum covariance determinant procedure searches for the subset h_i (out of n_i (sample size)) of the data set whose covariance matrix has the minimum determinant [23]. The sample observations based on the half set are chosen from the multivariate data set to obtain the *MCD* estimates of mean vectors and covariance matrices. These robust estimates are computed based on the clean data set selected by the half set. The *MCD* estimates are substituted into the conventional Fisher's equations, say Equations (1-2) to obtain the robust Fisher linear classification rule. Detailed description of this method is contained in [23]. The *MCD* approach requires the correction factor to obtain unbiased and consistent estimates if the data set comes from a multivariate normal distribution. The correction factor is used for the *FAST-MCD* algorithm to compute the *MCD* estimates. Detailed description and theorem to compute the concentration steps based on the half set of the *MCD* technique is contained in [24]. The allocation procedure for this method is the same as that of the Fisher linear classification analysis.

4 The Performance of Linear Classification Techniques

When the data set used in training the model is applied to validate the model, the classification performance is upward bias. In this case, the result is totally bias because the same data is used for both training and validation. An unbiased classification results can be obtained if the data set is splitted into two, say training and validation. Under this consideration, the training data is used to training the model while the validation set is used to validate the model. A stable and accurate classification results are obtained if the two categories of data set are reshuffled and replicated over a well defined Monte Carlo sample size.

5 Simulation

This simulation is designed to investigate the comparison between classification performance based on the mean of the optimal probability and the mean of the marginal probability. The aim is to compute the number of correct classification for each group, the overall mean probability and the marginal probability is compared in order to determine performance and to investigate if the marginal probabilities sum up to unity. In a general note, the mean of the optimal probability of correct classification only specifies the performance benchmark and satisfy the first axiom of probability. In this simulation, the data set is generated based on the contaminated normal model, say $40N_3(0,1)+10N_3(0.9,9),$ the meaning of this is that majority of the data set was drawn from the normal distribution $40N_3(0,1)$ and the remaining data set was generated from the contaminated normal portion $10N_3(0.9,9)$ respectively. The generated data set was added together and reshuffled and divided into 58 % (29) training and 42 % (21) validation. The result reported is based on 1000 replications. The marginal probabilities of correct classification and misclassification for the Fisher's technique are reported in Table 1.

Table 1. Probabilities of correct classification (optimal=0.8907)

Performance	Group one	Group two	Total
Correct classification	0.4524	0.4762	0.9286
Misclassification	0.0476	0.0238	0.0714
	0.5000	0.5000	1.0000

From the marginal probabilities both groups account for 0.5 sample observations which satisfies the axioms of probability, 93% of the sample observations were correctly classified while 7% were misclassified. The mean probability of correct classification (0.8765) and standard deviation (0.0277) has 0.9286 marginal probability of correct classification. The marginal probability revealed that 39 out of the 42 sample observations were correctly classified and 3 misclassified. Table 2 below contain the classification result for the robust Fisher's approach based on the minimum covariance determinant.

Table 2. Probability of correct classification (optimal=0.8907)

Performance	Group one	Group two	Total
Correct classification	0.4524	0.4524	0.9048
Misclassification	0.0476	0.0476	0.0952
	0.5000	0.5000	1.0000

The marginal probability indicates that 90% of the sample observations were correctly classified whereas about 10% were misclassified. The mean probability of correct classification (0.8668) and the standard deviation (0.0274) revealed that the conventional Fisher linear classification analysis is robust. The robust Fisher's approach account for 90% (38 out of 42) of the sample observation whereas the conventional approach account for 93% (39 out of 42) of the total sample observations for both groups.

6 Conclusion

The new procedure adopted revealed the proportion of sample observation correctly classified and misclassified, respectively. The approach is useful in determining the number of sample observations classified and can also be used to determine robustness. The limitation of this technique is that the standard deviations for the respect technique cannot be computed automatically. However, we have introduced a new procedure for analyzing the classification performance of the linear classification techniques based on two groups. The new technique of analyzing the performance also satisfies the two axioms of probability. We are investigating classification performance based on Type 1 and Type 2 errors. Indeed, performance analysis based on Type 1 and Type 2 errors with respect to probability of correct classification or misclassification only symbolizes the rejection of the null hypothesis or the acceptance of the alternate hypothesis.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936;7(2):179-188.

[2] Qiao Z, Zhou L, Huang JZ. Sparse linear discrfiminant analysis witth applications to high dimensional low sample size data. IAENG International Journal of Applied Mathematics. 2009;39(1):1-14.

[3] Kuhn M, Johnson K. Applied predictive modeling: Springer; 2013.

[4] Johnson RA, Wichern DW. Applied multivariate statistical analysis. Pearson Prentice Hall, Upper Saddle River; 2007.

[5] Martins TG. Reduced-rank discriminant analysis; 2013. Available: tgmstat.wordpress.com/2013/12/12/reduced-rank-discriminant-analysis.

[6] Timm NH. Applied multivariate analysis: Springer; 2002.

[7] Basak I. Robust M-estimation in discriminant analysis. The Indian Journal of Statistics. 1998;60(2):246-268.

[8] Devlin SJ, Gnanadesikan R, Kettenring JR. Robust estimation of dispersion matrices and principal components. Journal of the American Statistiscal Association. 1981;76(374):354-362.

[9] Filzmoser P, Hron K. Outlier detection for compositional data using robust methods. Mathematical Geosciences. 2008;40(3):233-248.

[10] Hubert M, Rousseeuw PJ, Van Aelst S. High breakdown robust multivariate methods. Statistical Science. 2008;23(1):92-119.

[11] Jin J, An J. Robust discriminant analysis and its application to Identify protein coding regions of rice genes. Mathematical Biosciences. 2011;232(2):96-100.

[12] Kim SJ, Magnani A, Boyd SP. Robust Fisher discriminant analysis. Advances in Neural Information Processing System. 2005;18:659-666.

[13] Pires AM, Branco JA. Generalization of Fisher's linear discriminant; 1996. Available: www.math.ist.utl.pt/~apires/PDF/APJB_RP96.pdf.

[14] Roelant E, Van Aelst S, Williems G. The minimum weighted covariance determinant estimator. Metrika. 2009;70(2):177-204.

[15] Wu G, Chen C, Yan X. Modified minimum covariance determinant estimator and its application to outlier detection of chemical process data. Journal of Applied Statistics. 2011;38(5):1007-1020.

[16] Linnet K. On the sensitivity of linear discriminant analysis to sampling variation and analytic errors. Computers and Biomedical Research. 1988;21(2):158-168.

[17] Zuo Y. Robust location and scatter estimators in multivariate analysis. Available: WSPC/Trim Size:9in x6in for Review, DOI: 10.1142/9781860948886_0021, 2005;0-31.

[18] Huber PJ. Robust estimation of a location parameter. Annals of Mathematical Statististics. 1964;35:73-101.

[19] Mallows CL. On some topics in robustness: Technical memorandum. Murray Hill, New Jersey: Bell Telephone Laboratories; 1975.

[20] Lopuhaa HP. On the relation between S estimators and M estimators of multivariate location and covariance. The Annals of Statistics. 1989;17:1662-1683.

[21] Rousseeuw PJ. Least median of square regression. Journal of the American Statistical Association. 1984;79:871-880.

[22] Multivariate estimators with high breakdown point. Mathematical Statistics and its Applications. 1984;283-297.

[23] Hubert M, Van Driessen K. Fast and robust discriminant analysis. Computational Statistics and Data Analysis. 2004;45:301-320.

[24] Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. Technometrics. 1999;41(3):212-223.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=729&id=6&aid=7053