Classification Rule for Small Samples: A Bootstrap Approach

M. M. Rahman[#1], M. M. Hossain[#2] and A. K. Majumder[#3]

[#1]Dept. of Statistics, Islamic University, Kushtia, Bangladesh, Phone no. +8801718698811
[#2]Dept. of Statistics, Islamic University, Kushtia, Bangladesh, Phone no. +8801716657066
[#3]Dept. of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh, Phone no. +8801711145041

ABSTRACT

In a recent year, classification is computer implemented and most popular data mining technique. Thus in this paper, we address the issue of classification errors over small samples and propose a new Bootstrap based approach for quantifying the level of classification errors. We investigate the performances of classification techniques and observed that, Bootstrap based classification techniques significantly reduce the classification errors than the usual techniques of small samples. Thus, this paper proposes to apply classification techniques under Bootstrap approach for classifying objects in case of small samples.

Key words: Fisher's Linear Classification, Quadratic Classification, Neural Network Classification, Bootstrap Approach, Small Sample.

Corresponding Author: M. M. Rahman

INTRODUCTION

Classification is perhaps the most familiar and most popular data mining technique (M. H. Dunham) [1]. Examples of classification applications include image and pattern recognition, medical diagnosis, loan approval, detecting faults in industry applications, and classifying financial market trends. Estimation and prediction may be viewed as types of classification. When someone estimates your age or guesses the number of marbles in a jar, these are actually classification problems. Before the use of current data mining techniques, classification was frequently performed by simply applying knowledge of the data. Statistical classification is a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items (referred to as traits, variables, characters, etc) and based on a training set of previously labeled items. We still do not have single classifier that can reliably outperform all others on a given data set. The accuracy of a particular parametric classifier on a given data set will clearly depend on the relationship between the classifier and the data (C. M. Van Der Walt and E. Barnard) [2]. By developing statistical classification methods we can asses the performance of the assignment rule, the relative sizes of the classes can be measured formally the differences between classes can also be tested (D. J. Hand) [3].

Classification techniques cannot usually provide an error-free method of assignment (R. A. Johnson and D. W. Wichern) [4]. This is because there may not be a clear distinction between the measured characteristics of the populations. A good classification procedure should result in few misclassifications. The classification techniques (Fisher's Linear Classification, Quadratic Classification, and Neural Network Classification) have been proposed as a solution to this type of problem when the size of samples is small.

Bootstrap methods are considered in the application of statistical process control because they can deal with unknown distributions and are easy to calculate using a personal computer. Bootstrap technique was invented by Bradley Efron [5] and further developed by Efron and Tibshirani [6]. "Bootstrap" means that one available sample gives rise to many others by re-sampling. Efron (1981, 1982) [7,8] developed Bootstrap with inferential purposes. Efron (1983) [9] and Efron and Tibshirani (1997) [10] studied the leave-one-out bootstrap methods for small sample classification. Their methods improved error estimation by bootstrap re-sampling with training set separated from test set. Their findings with small samples are encouraging. We are thus motivated to investigate whether Bootstrap re-sampling improves a simple and straightforward method of estimating misclassification error.

In our analysis, we investigate the performances of classification techniques and observed that, most of the techniques may not give few misclassifications under small samples. Small samples behave as large samples under Bootstrap approach. So, if the sample size is small, we apply classification techniques through the effect of Bootstrap approach and make a comparative study with the usual techniques of small samples. Here we also investigate that Bootstrap based classification techniques used in this analysis performs better than the usual techniques of small samples.

The rest of the paper is organized as: section 2 discusses the Bootstrap methods. Section 3 discusses the classification techniques considered in this study. Section 4 discusses the data sets used in this study. Results and discussion are presented in section 5. Finally, section 6 we achieve a conclusion about the paper.

BOOTSTRAP METHODS

The Bootstrap is a computer intensive re-sampling technique introduced by Efron [5] where theoretical statistics are difficult to obtain (J. G. Dias and J. K. Vermunt) [11]. The ability to do a lot of computation extremely fast has led to the use of techniques that provide "new" sets of data by re-sampling numbers generated from a single data set (B. Efron and R. J. Tibshirani) [6]. The bootstrap method is a powerful tool for estimating the sampling distribution of a given statistic. The bootstrap estimate of the sampling distribution is generally better than the normal approximation based on the central limit theorem (Bickel, P. J., and Freedman; Singh, K.) [12,13], even if the statistic is not standardized (Beran, R. J.; Liu, R., and Singh, K.) [14,15]. Let $X_1, X_2, ..., X_n$ be an *iid* sample following the distribution F with mean, μ and variance σ^2. The standard bootstrap procedure is to draw with replacement a random sample of size n from $X_1, X_2, ..., X_n$. Denote the bootstrap sample by $X_1^*, X_2^*, ..., X_n^*$ and denote their mean and standard deviation by $\bar{\bar{X}}_n^*$ and S_n^*. Suppose F_n indicate the empirical distribution of $X_1, X_2, ..., X_n$, then the sampling distribution of $\left(\bar{\bar{\mathrm{X}}}_{\mathrm{n}}^* - \bar{\bar{X}}_n\right)$ under F_n is the bootstrap approximation of the sampling distribution of $\left(\bar{\bar{X}}_n - \mu\right)$ under F. Its approximation error is shown to be negligible by the proposition derived by Bickel and Freedman [12] and Singh [13]. The bootstrap technique provides the mean $\hat{\theta}_B$ of all the bootstrap estimators

$\hat{\theta}_B = \frac{\sum_{i=1}^{B} \hat{\theta}_i}{B}$, where $\hat{\theta}_i$ is the estimate using the ith bootstrap sample and B is the number of bootstraps.

The bootstrap handles cases where a standard distribution cannot be assumed. The control limit is estimated by re-sampling the observed data to estimate the distribution of the observed variable. For many problems in statistics, we are interested in the distribution of values from a random sample of the population. If the underlying distribution from which the values are drawn is known, we can use developed theory to generate the sampling distribution. Efron [5] suggested the use of bootstrapping when there is little or no significant information about the underlying distribution. The idea behind the bootstrap is very simple, namely that (in the absence of any other information), the sample itself offers the best guide of the sampling distribution. By re-sampling with replacement from the original sample, we can create a bootstrap sample, and use the empirical distribution of our estimator in a large number of such bootstrapped samples to construct confidence intervals and tests for significance.

CLASSIFICATIION TECHNIQUES

Data mining is a process to mine and organize data in useful and coherent collections (J. Han and M. Kamber; B. A. Aski and H. A. Torshizi) [16,17]. The aim of data mining is description and prediction. There are many strategies in data mining which can be led to the prediction. One of them is classification. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data (M. H. Dunham) [1].

Fisher's Linear Classification

Fisher-LDA considers maximizing the following objective:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

where S_B is the "between classes scatter matrix" and S_W is the "within classes scatter matrix". Note that due to the fact that scatter matrices are proportional to the covariance matrices we could have defined J using covariance matrices the proportionality constant would have no effect on the solution (M. Welling) [18]. The scatter matrices are:

$$S_B = \sum_c (\mu_c - \overline{x})(\mu_c - \overline{x})^T$$

$$S_W = \sum_c \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T$$

where, $\overline{x}$ is the overall mean of the data cases. Often you will see that for 2 classes S_B is defined as $S'_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$. This is the scatter of class 1 with respect to the scatter of class 2 and hence corresponds to computing the scatter relative to a different vector. By using the general transformation rule for scatter matrices:

$$S_{\mu+v} = S_\mu + Nvv^T + 2Nv(\mu - \overline{x})^T$$

with $S_\mu = \sum_i (x_i - \mu)(x_i - \mu)^T$ we can deduce that the difference is only a constant shift not depending on any relative distances between points. A study concerned with this function maximally separates the two populations and used to classify new observations.

Quadratic Classification

A quadratic discriminant analysis is a general extension of a linear discriminant analysis that assumes the same variance-covariance matrix of different classes (K. M. Lee, T. J. Herrman, S. R. Bean, D. S. Jackson and J. Lingenfelser) [19]. The individual variance-covariance matrix of each class is used as a classification criterion in a quadratic discriminant analysis. Among several alternative classification rules used to discriminate among classes, the Bayes rule was used to compute the posterior probability to assign an observation x to a single class(G). According to this rule, given prior probabilities p_i and p_j, the observation x belongs to class G_i, if

$$P(x/G_i).p_i > P(x/G_j).p_j \quad for\ i \neq j$$

where, $P(x/G_i)$ and $P(x/G_j)$ are the probability densities. A quadratic discriminant assigns the observation x to class G_i when the discriminant score $D_i(x)$, a measure of the generalized squared distance between x and class G, is minimized.

$$D_i(x) = 0.5(x-\mu_i)' \sum\nolimits_i^{-1}(x-\mu_i) + 0.5\log(|\Sigma_i) - \log(p_i)$$

where, μ_i is the mean of class i, and Σ_i is the population variance-covariance matrix of class G_i. The posterior probability for each of the possible classifications is then obtained using the computed discriminant score $D_i(x)$. An observation x is assigned to the class with the largest posterior probability. In a linear discriminant analysis, the notation Σ_i is replaced with Σ due to the same variance-covariance matrix assumption

$$D_i(x) = 0.5(x-\mu_i)' \sum\nolimits^{-1}(x-\mu_i) - \log(p_i)$$

Neural Network Classification

Artificial neural networks are sometimes called semi-parametric or non-parametric technique; are aggregations of perceptions (S. Dreiseitl, L. O. Machado) [20]. For multi-layer feed forward networks, the output is

$$O_N = \frac{1}{1+e^{-(\beta O_H + \beta_0)}} \text{ and this output is again taken as } P(1|x,\beta,\beta_0,\alpha)$$

Here, O_H is a vector of perceptron outputs, each with its own α parameters (determined based on the data set, usually by maximum-likelihood estimation); these perceptrons are usually called hidden neurons. Due to the nonlinearity in these hidden neurons, the output O_N of an artificial neural network is a nonlinear function of the inputs.

DATA USED IN THIS STUDY

This study considers a secondary data set called Salmon data (Growth-Ring Diameters) contains one state from United States of America and Canada (Consider first 12 and very last 12 observations in this study). Each population consist two variables. Also considers a data set used for discriminating owners from non-owners of riding mowers. Both the data sets were collected from K. A. Jensen and B. Van Alen of the States of Alaska Department of Fish and Game. Data are given in the book entitled “Applied Multivariate Statistical Analysis”; 5th Edition, written by Richard A. Johnson and Dean W. Wichern [4]. A simulated data set also use in this study, one is generated from Uniform distribution over the range 0 and 1 and other one from F distribution of size 24.

RESULTS AND DISCUSSION

In this section, we apply three classification techniques namely Fisher's simple linear classification, Quadratic classification and Neural Network classification. We first apply these techniques to the original data set (initial sample) and also apply these techniques under Bootstrap approach. The results are given in the following subsequent tables.

Table1. Results for Salmon data (Growth-Ring Diameters)

		Apparent Error Rate (APER) in %		
		Fisher's Simple Linear Classification	Quadratic Classification	Neural Network Classification
Initial Sample (Without Bootstrap)		50.00	49.00	50.00
Bootstrap results	B (No. of Bootstrap)			
	50	8.75	7.58	18.50
	100	7.79	6.91	21.58
	500	7.13	6.36	22.65
	1000	7.02	6.23	22.27
	15000	6.93	6.21	22.00

From Table1, it is clear that, Fisher's Linear Classification, Quadratic classification and Neural Network classification gives the apparent error rate for initial data sets are 50%, 49% and 50% respectively, whereas under Bootstrap approach these classification techniques reduces this apparent error rate significantly and reaches to 6.93%, 6.21% and 22% respectively. Also, we observe that apparent error rate decreases if the number of Bootstrap increases but after a certain number of Bootstrap the difference is negligible.

Table2. Results for owners and non-owners of riding mowers

		Apparent Error Rate (APER) in %		
		Fisher's Simple Linear Classification	Quadratic Classification	Neural Network Classification
Initial Sample (Without Bootstrap)		16.67	16.67	41.67
Bootstrap results	B (No. of Bootstrap)			
	50	10.16	9.83	31.50
	100	10.00	9.37	32.50
	500	9.86	9.36	34.80
	1000	9.78	9.30	34.62
	15000	9.74	9.24	34.24

From Table2, it is clear that, Fisher's Linear Classification, Quadratic classification and Neural Network classification gives the apparent error rate for initial data sets are 16.67%, 16.67% and 41.67% respectively, whereas under Bootstrap approach these classification techniques reduces this apparent error rate significantly and reaches to 9.74%, 9.24% and 34.24% respectively. Also, we observe that apparent error rate decreases if the number of Bootstrap increases but after a certain number of Bootstrap the difference is negligible.

Table3. Results for simulated data

		Apparent Error Rate (APER) in %		
		Fisher's Simple Linear Classification	Quadratic Classification	Neural Network Classification
Initial Sample (Without Bootstrap)		25.00	16.70	25.00
Bootstrap results	B (No. of Bootstrap)			
	50	20.83	12.58	22.33
	100	20.33	12.75	21.83
	500	20.15	11.90	20.80
	1000	20.18	11.90	21.33
	15000	20.17	11.72	21.98

From Table3, it is clear that, Fisher's Linear Classification, Quadratic classification and Neural Network classification gives the apparent error rate for initial data sets are 25%, 16.70% and 257% respectively, whereas under Bootstrap approach these classification techniques reduces this apparent error rate significantly and reaches to 20.17%, 11.72% and 21.98% respectively. Also, we observe that apparent error rate decreases if the number of Bootstrap increases but after a certain number of Bootstrap the difference is negligible.

CONCLUSION

From our analysis, we investigate that, in case of small samples classification techniques significantly reduce the classification errors under Bootstrap approach. Fisher's Linear Classification gives the apparent error rate for initial data sets are 50%, 16.67% and 25% whereas, classification techniques under Bootstrap approach reduces this apparent error rate significantly at the level 6.93%, 9.74% and 20.17% respectively. We also investigate that, Quadratic Classification gives the apparent error rate for initial data sets are 49%, 16.67% and 16.70% whereas, it reduces these apparent error rate significantly under Bootstrap approach at the level 6.21%, 9.24% and 11.72% respectively, also Neural Network Classification gives the apparent error rate for initial data sets are 50%, 41.67% and 25% whereas, it reduces these apparent error rate significantly under Bootstrap approach at the level 22%, 34.24% and 21.98% respectively. It is clear from our analysis that, classification techniques under Bootstrap approach performs better than the usual techniques of small samples. "Thus we may conclude that, in case of small samples, propose to apply classification techniques under Bootstrap approach for classifying objects".

REFERENCES

[1] M. H. Dunham; Data Mining Introductory and Advanced Topics, Pearson Education (Singapor) Pte. Ltd., 2003.

[2] C. M. Van Der Walt and E. Barnard; Data Characteristics that Determine Classifier Performance, *South African Ins. of Electrical Eng.*, Vol. 98(3), pp. 87-93 Sept. 2007. *(Available online at http://www.saiee.org.za/publications/2007/Sept/98_3_3.pdf)*

[3] D. J. Hand; Discrimination and Classification, John Wiley and Sons, New York, 1981.

[4] R. A. Johnson and D. W. Wichern; Applied Multivariate Statistical Analysis, 5th ed., Pearson Education (Singapor) Pte. Ltd., 2002.

[5] Efron, B.; Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, Vol. 7, No. 1, 1979, pp. 1-26.

[6] B. Efron and R. J. Tibshirani; An Introduction to the Bootstrap, 1st ed., Chapman and Hall London, UK, 1993.

[7] Efron, B.; Nonparametric standard errors and confidence intervals, *Canadian Journal of Statistics*, Vol. 9, pp. 139-172, 1981.

[8] Efron, B.; The jackknife, the bootstrap and other re-sampling plans, *SIAM monographs*, No. 38, 1982.

[9] Efron, B.; Estimating the error rate of a prediction rule: Improvements on cross-validation, *J. Amer. Statist. Assoc.*, Vol. 78, pp. 316–331, 1983.

[10] Efron, B. and Tibshirani, R.; Improvements on cross-validation: the .632+ bootstrap method, *J. Am. Stat. Assoc.*, Vol. 92, pp. 548–560, 1997.

[11] J. G. Dias and J. K. Vermunt; Bootstrap Methods for Measuring Classification Uncertainty in Latent Class Analysis, 2006. *(Available online at http://www.academia.edu/2197720/Bootstrap_methods_for_measuring_classification_uncertainty_in_latent_class_analysis)*

[12] Bickel, P. J., and Freedman, D. A.; Some Asymptotic Theory for the Bootstrap, *The Annals of Statistics*, Vol. 9, pp. 1196-1217, 1981.

[13] Singh, K.; On the Asymptotic Accuracy of Efron's Bootstrap, *The Annals of Statistics*, Vol. 9, pp. 1187-1195, 1981.

[14] Beran, R. J.; Estimated Sampling Distributions: The Bootstrap and Competitors, *The Annals of Statistics*, Vol. 10, pp. 212-215, 1982.

[15] Liu, R., and Singh, K.; On a Partial Correction by the Bootstrap, *The Annals of Statistics*, Vol. 15, pp. 1713-1718, 1987.

[16] J. Han and M. Kamber; Data Mining: Concepts and Techniques, Elsevier Science and Technology Books, 2006.

[17] B. A. Aski and H. A. Torshizi; The Use of Classification Techniques to Enrich e-Learning Environments, *Online Proceedings of the University of Salford Fifth Education in a Changing Environment Conference Critical Voices*, pp. 135-39, Critical Times September, 2009. *(Available online at http://www.ece.salford.ac.uk/cms/resources/uploads/File/Paper%2012.pdf*)

[18] M. Welling; Fisher Linear Discriminant Analysis, *Dept. of Computer Science, University of Toronto, Canada, This is a note to explain Fisher Linear Discriminant Analysis. (Available online at http://www.cs.huji.ac.il/~csip/Fisher-LDA.pdf*)

[19] K. M. Lee, T. J. Herrman, S. R. Bean, D. S. Jackson and J. Lingenfelser; Classification of Dry-Milled Maize Grit Yield Groups Using Quadratic Discriminant Analysis and Decision Tree Algorithm, *AACC International Inc.*, Vol. 84, No. 2, pp. 152-61, 2007,
(Available online at http://naldc.nal.usda.gov/download/18571/PDF)

[20] S. Dreiseitl, L. O. Machado; Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review, *Journal of Biomedical Informatics*, Vol. 35, pp. 352-359, 2003. *(Available online at http://www.sciencedirect.com/science/article/pii/S1532046403000340)*