International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Performance of Classification Techniques along with Support Vector Machine

Muthukrishnan. R, Udaya Prakash. N

Abstract : Statistical learning is one of the most notable fields studied by the researchers to understand the data in the present scenario. Recent advances in the field of machine learning and artificial intelligence have been keen to develop more powerful automated techniques for predictive modeling, specifically in regression and classification models. These approaches fall under supervised statistical learning techniques, many conventional techniques are very complex to the data when it has larger volumes, i.e., if the data deviates from the model assumption, then the conventional procedure's results does not have the trustworthy. This paper explores and compares the classical methods with the alternatives in the context of classification, like logistic regression and support vector machine. The efficiency of these procedures has been evaluated through various measures such as confusion matrix and misclassification rate under real environment.

Keywords: Machine Learning - Logistic Regression – Support Vector Machine.

I. INTRODUCTION

In the information era, a vast amount of data is being generated in many fields. There is a need for valid statistical tools to extract the patterns, trends and to understand the data. Conventional procedures are reliable only when the data met the assumptions such as normality, linearity, and homogeneity. In this context, statistical learning is one of the notable fields identified by the researchers. It may be classified into supervised and unsupervised learning. The objective of supervised learning is to predict the outcome measure based on the input measures. The formation of supervised learning has been described as the practice of learning the relationship between precise inputs and outputs with the combination of the frequent criticism using the most powerful techniques applied for these kinds of learning are regression and classification. Many of the conventional and robust classification techniques rely on distributional assumptions such as multivariate normality or elliptical symmetry (Huber and Van Driessen, 2004). Most robust approaches use the concept of MCD and MVE estimators, which measure the centrality of a point relative to a multivariate sample. Literature shows that robust procedures outperforms over the conventional procedures, specifically when the data contains extreme observations, the data with violated assumptions. This paper has a choice of techniques used beneath of supervised learning such as linear discriminant analysis (LDA), robust discriminant analysis (RLDA), logistic regression (LOGIT) and Support Vector Machine (SVM) applied to improve the predictive models for an powerful automated techniques which will result in finding its effective and accurate in decision making. The rest of the paper is systematized to hold on its methodology of supervised learning techniques in section 2, followed by is experimental studies using two real datasets in section 3, finally the paper ends up with a conclusion in the last section.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Dr.R.Muthukrishnan*, Professor, Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu, India. E-mail: muthukrishnan1970@gmail.com
N.Udaya Prakash , Research Scholar, Department of Statistics, Bharathiar University, Coimbatore , Tamil Nadu, India. E-mail: udpk.06@gmail.com

II. SUPERVISED LEARNING PROCEDURES

Nowadays, vast amounts of data are being generated in many fields and that are corrupted by noise. It is a critical issue to develop supervised learning techniques that are immune to data uncertainties and perturbations. This section has the essentials on supervised learning which include the classical linear discriminant analysis, robust discriminant procedure, logistic regression and support vector machine.

A. Logistic Regression (LOGIT)

If the response variable is binary and there are many predictor independent variables that are either continuous or categorical. The main objective of binary logit model is classifying an observation into one of the two groups the concept is closely related to two-group discriminant analysis. Further, the model guarantees the probability of an observation belonging to a particular group between 0 and 1. Logistic regression in a predictive modeling technique allows identifying factors affecting various categorical outcomes of interest and the probability of those outcomes. The relationship between probability P and X_1, X_2,..., X_k is described by the following equation:

$$P = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_k x_k}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_k x_k}}$$

$$Log\left(\frac{P}{1-P}\right) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_k x_k$$

Jakramate and Ata (2012) discussed the classical problem of learning a classifier of logistic regression and multinomial logistic regression determined by a robust version.

B. Linear Discriminant Analysis (LDA)

Discriminant analysis is the appropriate tool for practice when the dependent variable is categorical (nominal) and the independent variables are interval or continuous.

The objective of discriminant analysis is to use the information from the independent variables to achieve the clearest possible separation or discrimination among groups.

The coefficients or weights (b), are estimated so that the groups differ as much as possible on the values of the discriminant function and this occurs when the ratio of the between-group sum of squares to the within-group sum of squares for the discriminant scores is at a maximum. The discriminant analysis model involves linear combinations of the following form:

$$D = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_k x_k$$

where D is the discriminant score, b's are Discriminant coefficients and X's are predictor or independent variable.

The conventional approach is mainly based on the sample mean vector and covariance matrix which is very sensitive to extreme observations. The result obtained by Lachenbruch (1975), says that discriminant analysis helps to estimate the error rates based on its variables selection when it is been considered for robustness.

C. Robust Linear Discriminant Analysis (RLDA)

The conventional LDA model can be very complex to the datasets with violated assumptions of the model and produces unreliable results. To rectify this issue one can apply the robust alternative, RLDA in order to make the dataset to proceed in a systematically improved one hence it will avoid the sensitivity in the given dataset by clearly integrating the model to classify it and optimize the worst case, because it follows the MCD (minimum covariance determinant (Rousseeuw (1985)) estimation method to estimate the measure of location and scatter matrix and this technique has greater influence to reduce the distance among the values, however, RLDA will perform well when compared with classical LDA. In real life problem results confirms RLDA models provide an equal performance or improved one that of LDA.

D. Support Vector Machine (SVM)

Support vector machines are one of the supervised statistical learning techniques (Vladimir (1995)). SVM is primarily used to classify data into different classes as done by some set of rules, which sort the help of the hyperplane that acts as an exceptional margin between the various classes.

It is also used to generate multiple separating hyperplanes such that the data is divided into segments and each segment contains only one kind of data. The SVMs is to be unaffected by potential noise on the input data. Here noise means that the values of data points might be influenced by measurement errors or may shift the data points in the input space.

If the data is linearly separable, then a pair *(w,b)* exists such that $w^T x_i + b \geq 1$, for all $x_i \in P$ and $w^T x_i + b \leq -1$, for all $x_i \in N$, with the objective is to find a hyperplane, $f_{w,b}(x) = sign(w^T x + b)$ that correctly classify our data. Here, $\boldsymbol{w}$ is the weight vector and $\boldsymbol{b}$ the bias (or $-\boldsymbol{b}$ is threshold).

An optimum separating hyperplane can be found by minimizing the squared norm of the separating hyperplane. Once the optimum hyperplane is found, data points, i.e support vectors that lie on its margin and the solution is a linear combination of only these points. After that, an appropriate the kernel function is used to map the new points into the feature space for classification. For multi-group classification the SVM method considered a set of binary classifications. The SVM is most for suitable for non linear relationship exits response and predictor variables and also multicollinearity is present in the dataset.

III. EXPERIMENTAL RESULTS

This section mainly focused on the performance of the supervised learning techniques in the context of classification problems. The two real data sets were considered for the study. (i) The hemophilia data (Habemma et al. (1974)) contains two measured variables AHF activity and AHV antigen on 75 women, belonging to two groups as the first group contains 30 observations belong to normal group and the second group contains 45 observations that belong to obligatory carrier. (ii) The anorexia data (Hand et al. 1993) contains 3 groups, each group two variables with a frame of 72 observations. The weight change data for young female anorexia patients. The two variables are, prewt (weight of patients before study periods) and postwt (weight of patients after study periods), classified the three groups, namely CBT (Cognitive-behavioral treatment), Cont (Control), FT (Family treatment).

Classification analysis was performed for these data sets under with and without outliers by using various classification procedures. The outliers were identified via distance-distance plots (Figure 1 and is given in the appendix). The obtained results are summarized and are given in the form of the classification matrix (Table 1) and misclassification probabilities (Table 2) under various classification procedures. The results show that the SVM classification procedure shows more classification accuracy followed by RLDA when compared with logistic and LDA.

Table 1: Classification matrix under various classification procedures

Dataset	No. of observations	LOGIT	LDA	RLDA	SVM
Hemophilia	$\begin{bmatrix}45 & 0\\ 0 & 30\end{bmatrix}$	$\begin{bmatrix}41 & 4\\ 5 & 25\end{bmatrix}$	$\begin{bmatrix}38 & 7\\ 4 & 26\end{bmatrix}$	$\begin{bmatrix}38 & 7\\ 4 & 26\end{bmatrix}$	$\begin{bmatrix}41 & 4\\ 4 & 26\end{bmatrix}$
	$\begin{bmatrix}\mathbf{45} & \mathbf{0}\\ \mathbf{0} & \mathbf{25}\end{bmatrix}$	$\begin{bmatrix}\mathbf{41} & \mathbf{4}\\ \mathbf{4} & \mathbf{21}\end{bmatrix}$	$\begin{bmatrix}\mathbf{39} & \mathbf{6}\\ \mathbf{4} & \mathbf{21}\end{bmatrix}$	$\begin{bmatrix}\mathbf{39} & \mathbf{6}\\ \mathbf{4} & \mathbf{21}\end{bmatrix}$	$\begin{bmatrix}\mathbf{41} & \mathbf{4}\\ \mathbf{3} & \mathbf{22}\end{bmatrix}$
Anorexia	$\begin{bmatrix}29 & 0 & 0\\ 0 & 26 & 0\\ 0 & 0 & 17\end{bmatrix}$	$\begin{bmatrix}12 & 10 & 7\\ 9 & 17 & 0\\ 6 & 4 & 7\end{bmatrix}$	$\begin{bmatrix}11 & 10 & 8\\ 9 & 17 & 0\\ 6 & 4 & 7\end{bmatrix}$	$\begin{bmatrix}12 & 9 & 8\\ 9 & 17 & 0\\ 6 & 4 & 7\end{bmatrix}$	$\begin{bmatrix}22 & 5 & 2\\ 6 & 19 & 1\\ 5 & 3 & 9\end{bmatrix}$
	$\begin{bmatrix}\mathbf{18} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{26} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{13}\end{bmatrix}$	$\begin{bmatrix}6 & 12 & 0\\ 5 & 21 & 0\\ 0 & 0 & 13\end{bmatrix}$	$\begin{bmatrix}\mathbf{5} & \mathbf{13} & \mathbf{0}\\ \mathbf{3} & \mathbf{22} & \mathbf{1}\\ \mathbf{0} & \mathbf{0} & \mathbf{13}\end{bmatrix}$	$\begin{bmatrix}\mathbf{5} & \mathbf{13} & \mathbf{0}\\ \mathbf{2} & \mathbf{23} & \mathbf{1}\\ \mathbf{0} & \mathbf{0} & \mathbf{13}\end{bmatrix}$	$\begin{bmatrix}\mathbf{16} & \mathbf{2} & \mathbf{0}\\ \mathbf{6} & \mathbf{19} & \mathbf{1}\\ \mathbf{0} & \mathbf{0} & \mathbf{13}\end{bmatrix}$

[.] – Without outliers

Table 2: Misclassification probabilities under various classification procedures

Datasets	Logit	LDA	RLDA	SVM
Hemophilia - 2 Groups **(70)**	0.120 **(0.114)**	0.147 **(0.143)**	0.147 **(0.143)**	**0.107** **(0.100)**
Anorexia - 3 Groups **(57)**	0. 500 **(0.299)**	0.514 **(0.299)**	0.500 **(0.281)**	**0.306** **(0.158)**

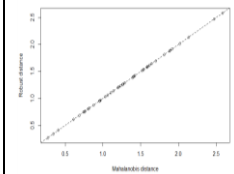
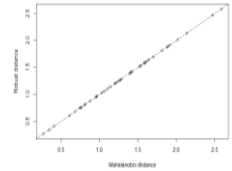
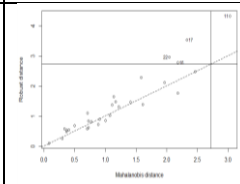
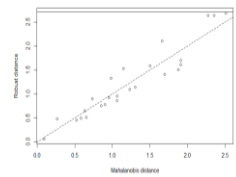
(.) - Without outliers

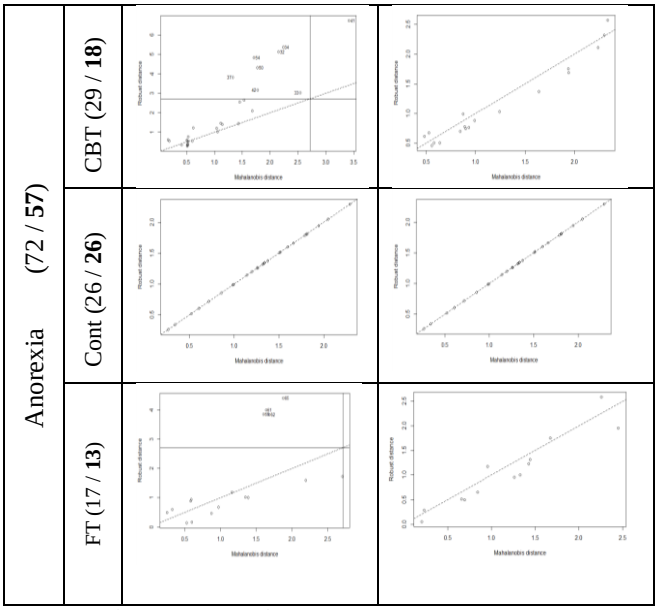
The above table shows the superiority of the SVM procedure since the SVM model gives very fewer misclassification probabilities when compared with the other classification procedures.

IV. CONCLUSION

The conventional procedures should perform reasonably well if certain assumptions hold but may be unreliable if one or more of these assumptions are violated. The sample means vector and covariance matrix is very sensitive to outliers. Hence the conventional LDA does not provide reliable results when the data contains outliers. The least-squares approach is also very sensitive to extremes observations. Hence the results produced by Logistic regression also unreliable when the data deviate from the model assumptions. There is a need for a robust alternative to increase accuracy even when the data slightly deviate from the model assumptions for non-normal situations. RLDA performs well next to SVM when compared with LOGIT and LDA. The study concluded that classification with SVM gives more accuracy followed by RLDA and then other procedures. To increase the accuracy further, the study may be extended by applying suitable kernel/robust procedures in computational aspects of the SVM algorithm

APPENDIX

Dataset		With outliers	Without outliers
Hemophilia (75 / **70**)	Carrier (45 / **45**)		
	Normal (30 / **25**)		



(.) – no. of cases; **(.) - no. of cases without outliers**

Figure 1: Distance-Distance Plots

REFERENCES

1. Boser, B.E.,Guyon, I.M., Vladimir N.V. A training algorithm for optimal margin classifiers, Proceedings of the fifth annual workshop on Computational learning theory,1992 144.
2. Cortes, C. and Vladimir, N.V. Support-vector networks. Machine Learning. 20,1995, 273–297.
3. Cramer, J. S. The origins of logistic regression. Tinbergen Institute. 2002,167–178.
4. Crowley, M.J., The R Book, John Wiley and Sons, Limited,2007.
5. Jakramate, B and Ata, K. , Label-noise Robust Logistic Regression and Its Applications. European Conference on Machine Learning and Knowledge Discovery in Database.2012, 143 – 158.
6. James.G., Witten.D., Hastie,T., and Tibshirani,R. An Introduction to Statistical Learning with applications in R, Springer,2017.
7. John, F.Applied regression analysis, linear models, and related models, Sage publications, Inc,1997.
8. Lachenbruch, P. A. Discriminant analysis. NewYork, Hafner,1975.
9. Leys, C., Klein, O., and Dominicy, Y., Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. Journal of Experimental Social Psychology, 74, 2018, 150–156.
10. Mahalanobis, P. C., On the generalised distance in statistics. Proceedings of the National Institute of Sciences, 12, 1936, 49–55.

11. Muthukrishnan, R and Mahesh, K., Evaluation of classical and robust Discriminant Methods under Apparent Error Rate, International Journal of Current Research, 5, 2013, 2817-2820.
12. R Core Team (2019). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. URL, http://www.r-project.org.
13. Rousseeuw, P.J., Multivariate estimation with high breakdown point, Mathematical Statistics and Applications, 8, 1985, 283-297.
14. Samuel A.L., Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. 3, 1959, 210–229.
15. Vladimir, N.V., The Nature of Statistical Learning Theory, Springer, 1995.

AUTHORS PROFILE

Dr. R. Muthukrishnan, was born in Tirunelveli. He holds B.Sc. (1991) in Computer Science from Madurai Kamaraj University, M.Sc. (1993) and Ph.D. (2000) in Statistics from Manonmaniam Sundaranar University. He is currently working as a Professor in the Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu, INDIA. He is a life member of various academic bodies such as ISPS, IBS, IISA and member in ISI and ISA. His main research work focuses on Robust Statistical Inference and Multivariate Analysis. He has more than 18 years of teaching and research experience. He has published 50 research papers in reputed national/international journals and guided 26 scholars for their research programs.

N.Udaya Prakash, Research scholar, Department of Statistics, Bharathiar University, His research area is Statistical Inference predominantly in robust Inference techniques with SVM procedures using R software.