Research Article

EVALUATION OF LOGISTIC REGRESSION MODEL WITH FEATURE SELECTION METHODS ON MEDICAL DATASET

[1]Raghavendra B. K., [2]Dr. Jay B. Simha

Address for Correspondence

[1]Dr. M.G.R. Educational and Research Institute, Chennai-600 095

[2]Abiba Systems, Bengaluru-560 050

Email: raghavendra_bk@rediffmail.com, jay.b.simha@abibasystems.com

ABSTRACT

Logistic regression is a well known classification method in the field of statistical learning. It allows probabilistic classification and shows promising results on several benchmark problems. Logistic regression enables us to investigate the relationship between a categorical outcome and a set of explanatory variables. The outcome or response can be either dichotomous (yes, no) or ordinal (low, medium, high). During dichotomous response, we are performing standard logistic regression and for ordinal response, we are fitting a proportional odds model. In this research work an attempt has been made to introduce model that uses standard logistic regression formula with feature selection using forward selection and backward elimination methods and has been evaluated for the effectiveness of the results on publicly available medical datasets. The process of evaluation is as follows. The feature selection algorithm using forward selection and backward elimination method is applied on the dataset and the selected features from these algorithms are used to develop a predictive model for classification using logistic regression. The classification accuracy, root mean square error, and mean absolute error are used to measure the performance of the predictive model. From the experimental results it is observed that logistic regression model with feature selection using forward selection and backward elimination methods gives more reliable result than the logistic regression model.

KEYWORDS Backward elimination, dichotomous variable, explanatory variable, feature selection, forward selection, logistic regression, medical dataset.

I. INTRODUCTION

In the last few years, digital revolution has provided relatively inexpensive and available means to collect and store large amounts of patient data in database, i.e., containing rich medical information and made available through the Internet for Health Services globally. Data mining techniques logistic regression is applied on these databases to identify the patterns that are helpful in predicting or diagnosing the diseases and to take therapeutic measure of those diseases.

Logistic regression is a technique for analyzing problems in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). In logistic regression, the dependent variable is binary or dichotomous, i.e., it only contains data coded as 1 (TRUE, success, etc.) or 0 (FALSE, failure, etc.).

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients of formula to predict a logit transformation for the probability of presence of characteristic of interest.

The rest of the paper is organized as follows: Section 2 reviews the prior literature, Logistic regression technique is discussed in Section 3. Experimental validation using publicly available medical dataset is given in Section 4. Section 5 includes Experimental results and discussions followed by conclusion.

II. LITERATURE SURVEY

There is an approach that examines the problem of efficient feature evaluation for logistic regression on very large data sets. The authors present a new forward feature selection heuristic that ranks features by their estimated effect on the resulting model's performance. An approximate optimization, based on back fitting, provides a fast and accurate estimate of each new feature's coefficient in the logistic regression model. Further, the algorithm is highly scalable by parallelizing

simultaneously over both features and records, allowing us to quickly evaluate billions of potential features even for very large data sets [3].

Recent studies of machine learning algorithms in high-dimensional data revealed that the three top performing classes of algorithms for high-dimensional data sets are logistic regression, Random Forests and SVMs [4]. Although logistic regression can be inferior to non-linear algorithms, e.g. kernel SVMs, for low-dimensional data sets, it often performs equally well in high-dimensions, when the number of features goes over 10000, because most data sets become linearly separable when the numbers of features become very large. Given the fact that logistic regression is often faster to train than more complex models like Random Forests and SVMs, in many situations it is the preferable method to deal with high dimensional data sets [5].

However, even with a scalable algorithm it can still be computationally infeasible to use the billions of features that could be potentially useful. The choice of features in high dimensions can have a significant effect on the performance of the learned model and the computational tractability of the learning algorithm. Many algorithm-independent high level feature selection techniques are exist, however, in most cases the running time becomes an issue for large numbers of features. Although popular and extremely well established in mainstream statistical data analysis, logistic regression is strangely absent in the field of data mining. This article introduces two possible explanations of this phenomenon. First, there might be an assumption that any tool which can only produce linear classification boundaries is likely to be trumped by more modern nonlinear tools. Second, there is a legitimate fear that logistic regression cannot practically scale up to the massive dataset sizes to which modern data mining tools are applied. This article consists of an empirical examination of the first assumption, and surveys, implements and compares techniques by which logistic regression can be scaled to data with millions of attributes and records. The results, on a large life science dataset, indicate that logistic regression can perform surprisingly well, both statistically and computationally, when compared with an array of more recent classification algorithms [6].

Feature selection is a key task in remote sensing data processing, particularly in case of classification from hyper spectral images. A logistic regression (LR) model can be used to predict the probabilities of the classes on the basis of input features, after ranking them according to their relative importance. In this work, the LR model is applied for both feature selection and the classification of remotely sensed images, where more informative soft classifications are produced naturally. The results indicate that, with fewer restrictive assumptions, the LR model is able to reduce the features substantially without any significant decrease in the classification accuracy of both the soft and hard classifications [7].

Multivariate logistic regression is often used within the field of epidemiology to describe the relationship between disease occurrence and an exposure suspected to be associated with the disease [8]. Additional effects are added to the model if they confound the disease-exposure relationship. Traditional model selection procedures, which focus on selecting models that are good predictors of the dependent variables, are not necessarily the most appropriate for epidemiological research questions. The proposed backwards-manual selection macro, *%bms*, attempts to select logistic regression models more suitable for epidemiological

research. The macro consists of two main stages, (1) backwards selection of effect-modifiers, and (2) selection of main effects based on their confounding potential and influence on overall model-fit. During the first stage, the macro generates all first-order effect modifiers for the variables provided by the user and PROC Logistics' backwards selection option is used to remove non-significant effect-modifiers. The second stage begins by removing the least significant potential confounder from the model. If this does not cause a change in the relationship between disease and exposure or the overall model fit, it remains out of the model. This process continues until all of the potential confounders not included in an effect-modifier have been evaluated.

III. LOGISTIC REGRESSION

Logistic regression is also called as logistic model or logit model, is a type of predictive model which can be used, when the target variable is a categorical variable with two categories - for example live or die, has disease or doesn't have disease, purchase product or doesn't purchase product, wins race or doesn't win etc. Logistic regression is used for the prediction of the probability of occurrence of an event by fitting the data into a logistic curve. Like many forms of regression analysis, it makes use of predictor variables; variables may be either numerical or categorical. For example, the probability that a person has a heart attack within a specified time that might be predicted from the knowledge of person's age, sex and body mass index. Logistic regression is used extensively in the medical and social sciences as well as in marketing applications such as prediction of customer's propensity to purchase a product or cease a subscription.

The response, Y, of a subject can take one of two possible values, denoted by 1 and 0 (for example, Y=1 if a disease is present; otherwise Y=0). Let $X=(x_1, x_2, \ldots, x_n)$ be the vector of explanatory variables. The logistic regression model is used to explain the effects of the explanatory variables in the form of binary response.

$$\text{Logit}\ \{\Pr(Y=1|x)\} = \log\{\Pr(Y=1|x)\}/1-\Pr(Y=1|x) = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\ldots+\beta_kx_k \quad (1)$$

Where β0 is called the "intercept" and β_1, β_2, β_3, and so on are called the "regression coefficients" of x_1, x_2, x_3 respectively. Each of the regression coefficients describes the size of the contribution of the risk factor. A positive regression coefficient means that the risk factor increases the probability of the outcome, while a negative regression coefficient means that the risk factor decreases the probability of that outcome, a large regression coefficient means that the risk factor strongly influences the probability of that outcome, while a non-zero regression coefficient means that the risk factor has little influence on the probability of that outcome.

The logistic function is given by

$$P=1/(1+e^{-\text{logit}(p)}) \quad (2)$$

A graph of the function is shown in Fig. 1 below. The logistic function is useful because it can take an input any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1.

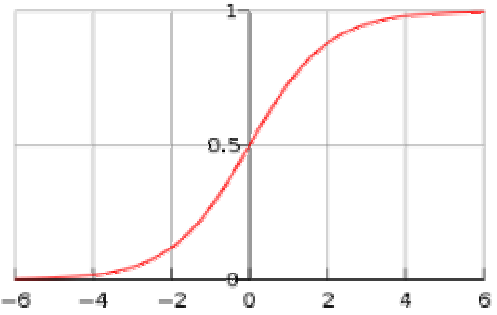

Fig. 1. A graph of logistic regression function

IV. EXPERIMENTAL VALIDATION

The process of evaluation is as follows. The feature selection algorithm using forward selection and backward elimination method is applied on the publicly available medical data set and the selected features from these algorithms are used to develop a predictive model for classification using logistic regression. For

evaluation ten fold cross validation has been used. In 10-fold cross validation, the original sample is partitioned into 10 sub samples, out of 10 sub samples, a single sub sample is taken as the validation data for testing the model, and the remaining sub samples are used as training data. The cross-validation process is repeated 10 times (the folds), with each of the 10 sub samples used exactly once as the validation data. The 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub sampling is that all observations are used for both training and validation and each observation is used for validation exactly once.

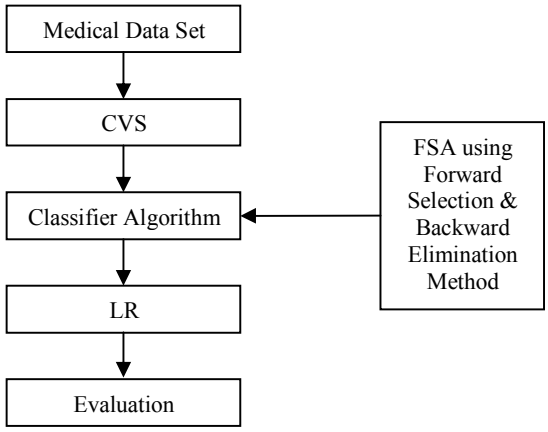

Fig. 2. Logistic Regression with Feature Selection Frame work

V. RESULTS AND DISCUSSION

The publicly available medical data sets are used in this work. In the evaluation forward search and backward elimination feature selection methods has been evaluated for the effectiveness of the classification using logistic regression.

The Classification Accuracy, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used to measure the performance of Logistic regression with feature selection model. Table 1 gives specifications for the datasets. The results of the evaluation using forward selection and backward elimination are given in Table 2 and Table 3 respectively. From the result it can be observed that the classification accuracy, RMSE, and MAE are more efficient when compared Logistic Regression Model and Logistic Regression with Feature Selection Model. The classification accuracy, RMSE, and MAE are almost similar. Fig. 3 and Fig. 4 gives the classification accuracy details after evaluation process for both forward selection and backward elimination methods.

Table 1: Specification for the medical dataset

Sl No	Medical Dataset	No of Instances	Total No of Attributes	No of Classes
1	Pima Diabetes	768	9	2
2	Hepatitis	155	20	2
3	Heart-c	303	14	5
4	Heart-h	294	14	5
5	Statlog-Heart	270	14	2
6	Bupa Liver Disorders	345	7	2
7	Spect Test	187	23	2
8	Wiscosin Breast Cancer	286	10	2
9	Haberman	306	4	2
10	Postoperative Patient Data	90	9	3

Table 2: LR and LR with feature selection using forward selection method specification on medical datasets

Sl No	Medical Dataset	Logistic Regression			Logistic Regression with Feature Selection using Forward Selection Method		
		CA	RMSE	MAE	CA	RME	MAE
1	Pima Diabetes	78.25	0.39	0.3	76.82	0.4	0.32
2	Hepatitis	89.03	0.26	0.14	88.38	0.29	0.17
3	Heart-c	87.12	0.19	0.07	87.12	0.19	0.07
4	Heart-h	87.07	0.18	0.07	87.07	0.18	0.07
5	Statlog-Heart	85.55	0.32	0.2	85.55	0.32	0.2
6	Bupa Liver Disorders	70.43	0.45	0.41	70.43	0.45	0.41
7	Spect Test	76.46	0.4	0.32	77	0.4	0.32
8	Wiscosin Breast Cancer	76.22	0.4	0.32	76.22	0.4	0.32
9	Haberman	75.49	0.41	0.33	75.49	0.41	0.33
10	Postoperative Patient Data	74.44	0.35	0.24	74.44	0.35	0.24

Table 3: LR and LR with feature selection using backward elimination method specification on medical datasets

Sl No	Medical Dataset	Logistic Regression			Logistic Regression with Feature Selection using Backward Elimination Method		
		CA	RMSE	MAE	CA	RME	MAE
1	Pima Diabetes	78.25	0.39	0.3	76.82	0.4	0.32
2	Hepatitis	89.03	0.26	0.14	89.67	0.28	0.16
3	Heart-c	87.12	0.19	0.07	87.45	0.2	0.08
4	Heart-h	87.07	0.18	0.07	86.73	0.19	0.07
5	Statlog-Heart	85.55	0.32	0.2	85.92	0.34	0.24
6	Bupa Liver Disorders	70.43	0.45	0.41	70.43	0.45	0.41
7	Spect Test	76.46	0.4	0.32	78.07	0.42	0.36
8	Wiscosin Breast Cancer	76.22	0.4	0.32	77.27	0.41	0.33
9	Haberman	75.49	0.41	0.33	75.81	0.41	0.34
10	Postoperative Patient Data	74.44	0.35	0.24	74.44	0.35	0.25

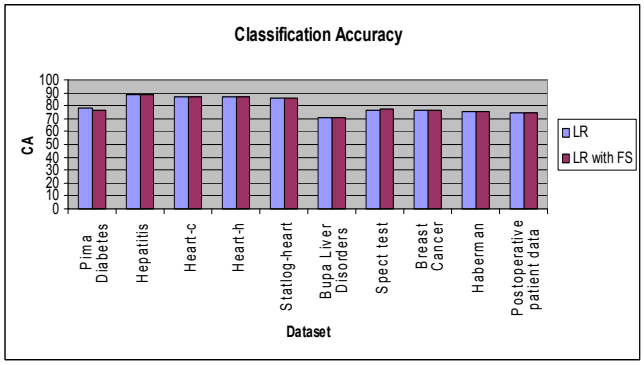

Fig.3: Classification results after evaluation (Forward Selection Method)

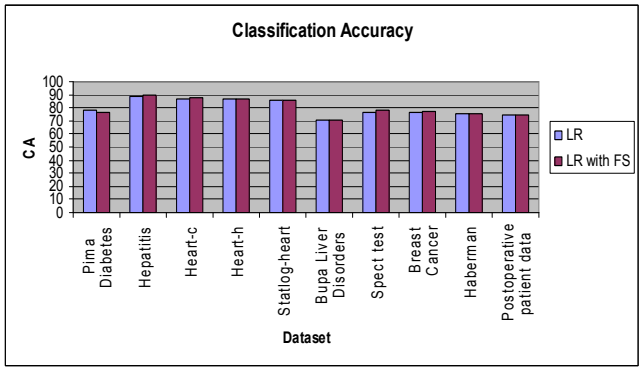

Fig. 4. Classification results after evaluation (Backward Elimination Method)

VI. CONCLUSION

In this research work an attempt has been made to evaluate Logistic Regression with Feature Selection Model on publicly available medical data sets and the selected features from algorithm are used to develop a predictive model for classification using logistic regression. The results have been conclusive about the effectiveness of logistic regression with feature selection methods and validate the hypothesis of the research.

REFERENCES

1. C.L. Blake, C.J Merz., "UCI repository of machine learning databases". [http://www.ics.uci. edu/~mlearn/ MLRepository.html], Department of Information and Computer Science, University of California, Irvine.
2. http://en.wikipedia.org/wiki/Logistic_regression
3. Singh, S., Kubica J., Larsen S., Sorokina D, "Parallel Large Scale Feature Selection for Logistic Regression", SIAM International Conference on Data Mining (SDM), 2009.
4. R. Caruana, N. Karampatziakis, and A. Yessenalina. "An empirical evaluation of supervised learning in high dimensions". In Proceedings of the 25th International Conference on Machine Learning, (ICML 2008), 2008.
5. P. Komarek and A. Moore. "Making logistic regression a core data mining tool with trirls". In Proceedings of the 5th International Conference on Data Mining Machine Learning, page 4, 2005.
6. Paul R. Komarek, Andrew W. Moore, "Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs", In Artificial Intelligence and Statistics (2003).
7. Qi Cheng Varshney, P.K. Arora, M.K., "Logistic Regression for Feature Selection and Soft Classification of Remote Sensing Data", Geoscience and Remote Sensing Letters, IEEE, vol. 3, pp 491-494, 2006.
8. Janice hegewald, Annette Pfahlberg, and Wolfgang Uter, "A Backwards-Manual Selection Macro for Binary Logistic Regression in the SAS v8.02 PROC LOGSTIC Procedure", NESUG 2003.
9. Piramuthu S., "Evaluation Feature Selection Methods for Learning in Data Mining Applications", 31st Hawaii International Conference on System Sciences, vol 5, pages 294-301, Jan 1998.
10. P.S.Yu C.C. Aggrawal. "Data mining techniques for associations, clustering and classification". In Proceedings of the 3rd Pacific-Asia Conference on

Methodologies for Knowledge Discovery and Data Mining, pp 13-23, 1999.
11. http://www.cs.waikato.ac.nz/ml/weka
12. Liu H., Sentino R, "Some issues on scalable Feature Selection, Expert Systems with Application", vol 15, pp 333-339, 1998.

Biographies and Photographs

Dr. Jay B. Simha is working as a Chief Technology Officer at Abiba Systems, Bangalore, INDIA. He has received M.Tech degree from Mysore University in Maintenance Engineering (1994), M.Phil. Degree in Computer science from MS University (2003) and PhD degree from Bangalore University in data mining and decision support (2004). He worked with Alpha Systems, Siemens and other companies in software development and analysis. His research areas are Data Mining, Fuzzy Logic, and Artificial Intelligence. He has published over 25 papers in refereed journals and conferences.

Raghavendra B K is working as a Assistant Professor & Head, Department of Computer Science and Engineering at Ghousia College of Engineering, Ramanagaram, Karnataka, India. He has received his BE degree in Computer Science & Engineering from Bangalore University (1994) and M.Tech degree in Computer Science & Engineering from Visveswaraya Technological University, Belgaum (2004), and pursuing his PhD at Dr. MGR University, Chennai, India. His research area is Data Mining.