International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161
©2014 INPRESSCO®, All Rights Reserved
Available at http://inpressco.com/category/ijcet

Research Article

Logistic Regression in Data Mining and its Application in Identification of Disease

Dhaval Sanghavi[Ȧ*], Hitarth Patel[Ȧ] and Sindhu Nair[Ȧ]

[Ȧ]Computer Science, Dwarkadas J.Sanghvi College of Engineering, Vile Parle(W),Mumbai-400056,India

Accepted 05 Nov 2014, Available online 01 Dec 2014, Vol.4, No.6 (Dec 2014)

Abstract

Data mining in clinical medicine deals with learning models to predict health of patients. The models is used to support clinicians in therapeutic or monitoring tasks. Data mining techniques are usually applied in clinical contexts to analyze retrospective data, thus giving professionals to check large amounts of data routinely collected during their day activity. Moreover, clinicians can take advantage of data mining techniques to deal with the amount of research results obtained by molecular medicine, which may allow transition from population-based to personalized medicine.Logistic regression is used to analyze relationships between a dichotomous dependent variable and metric or dichotomous independent variables.

Keywords: *logistic regression, feature extraction, Data Mining*

1. Introduction

Predictions may range from the simple stratification of the patients' population on the basis of known risk factors, such as age or lifestyle, to the forecast of the effect that a treatment or drug may have on a single patient. Generally speaking, in a clinical context, predictions may support diagnostic, therapeutic, or monitoring tasks. Diagnosis is related to the classification of patients into disease classes or subclasses on the basis of patients' data.

Logistic regression is used to analyze relationships between a dichotomous dependent variable and metric or dichotomous independent variables. Logistic regression combines the independent variables to estimate the probability that a event will occur . The variate or value produced by logistic regression is a probability value between 0.0 and 1.0. If the probability for group membership in the modeled category is above some cut point (the default is half), the subject is predicted to be a member of the modeled group. If the probability is less than the cut point, the subject is included to be a member of the other group. For any given case, logistic regression computes the probability that a case with a particular set of values for the independent variable is a member of the modeled category.

Yi=e^u/(1+e^u)

where Yi is define as the estimated probability that the ith case is in a category and u is the regular linear regression equation:

Feature selection based data mining methods is one of the most important research directions in the fields of machine learning. Especially in recent years, along with the appearance of many high dimension / small sample problems, such as, natural language processing, biological information, economic and financial, network and telecom, and medical data analysis, the study of feature selection once again become the focus of attention. Identifying biomarkers with high sensitivity and specificity for high mortality and morbidity diseases such as UA etc. plays a key role in diagnosis and prognosis for them. Moreover, recent years have seen increasing research interest in biomarker identification is turned from one specific biomarker to biomarker pattern with interactions. We had found that feature selection based data mining methods better fit to investigate syndrome of biological basis . The definition of “Characteristic pattern” was proposed to reduce the gap between “golden index” and biological basis of syndrome .

2. Material and Methods

2.1 Statistical methods to detect metabolites and proteins with significant change Independent sample t test and analysis of variance

ANOVA was used to detect biomarkers with significant change of concentrations between the disease and control samples. P value was calculated to measure significance of mean and variance of each biomarker between the two groups. It is noted that a sample with missing value was not included to calculating p value since arithmetic mean may obliterate the significance of a biomarker.

2.2 Feature selection based data mining methods

We use three kinds of feature selection methods: Filter, Wrapper and Embedded, to carry out a comparison study

*Corresponding author: **Dhaval Sanghavi** to select biomarkers for the disease. Furthermore, interactions of biomarkers were studied by integrating domain knowledge into feature selection models built via the three data mining methods. k-Fold cross validation was used to evaluate performance of three supervised regression methods. Sensitivity, Specificity and Accuracy were computed and compared for each method to determine a better data mining method.

2.3 Detecting biomarkers with significant difference by t test in statistics

Independent t test was done by SPSS (version 17.0) to detect significantly different metabolites and proteins between UA and control patients. $P < 0.05$ was considered as of significance.

3. Results

3.1 Comparison study of feature selection methods

We employed three hackneyed performance measures: accuracy, sensitivity and specificity. A distinguished confusion matrix is obtained to calculate the three methods. Confusion matrix is a matrix representation of the classification results. the upper left denotes the number of samples classifies as true while they were true (i.e., true positives), and lower right cell denotes the number of samples classified as false while they were actually false (i.e., true false). The other two cells (lower left cell and upper right cell) denote the number of samples wrongly classified. Specifically, the lower left cell denoting the number of samples classified as false while they actually were true (i.e., false negatives), and the upper right cell denoting the number of samples classified as true while they actually were false (i.e., false positives).Once the confusion matrixes were constructed, the sensitivity, accuracy, sensitivity and specificity are calculated as: sensitivity = TP/(TP + FN); specificity = TN/(TN + FP). Accuracy = (TP + TN)/(TP + FP + TN + FN); where TP, TN, FP and FN denotes true positives(TP), true negatives(TN), false positives(FP) and false negatives(FN), respectively.

Table 1: Comparison study of three feature selection methods in metabolomics data

Methods	Number of metabolites	Classification Accuracy
Filter	4	73.81%
Wrapper	1	71.43%
Embedded	3	71.43%

Table 2: Comparison study of three feature selection methods in proteomics data

Methods	Number of proteins	Classification Accuracy
Filter	11	76%
Wrapper	4	72%
Embedded	2	76%

From Table 1 and Table 2, it is found that Filter performs better than the other two feature selection methods in metabolomics while Embedded is better in selecting as few as proteins with maximal classification accuracy. Since C4.5 decision tree was chosen as embedded feature selection method here to compare with the other methods, it can be clearly plotted by using concentration of selected metabolites or proteins.

3.2 A novel feature selection method to choose metabolite and protein biomarkers for UA

We used the variance analysis to choose several proteins with possible significant difference between UA and control and then rank the included metabolites and proteins by F value calculated for them. The classification model was chosen as neural network we used a popular ANN architecture called multi-layer perceptron (MLP) with back-propagation. The MLP is known to be a powerful function approximator for prediction . It is arguably the most commonly used and well studied ANN architecture. So many literatures have reported its good performance in performing classification task

Conclusion

In this paper, we compared the three types of feature selection based data mining methods, and presented a novel computational strategy to select biomarkers as few as possible for disease. The results show that Wrapper based feature selection methods perform best than Filter and Embedded methods. Based on this, we presented a novel feature selection by combing t test and classification data mining method to select 6 proteins with 96% prediction accuracy and 5 metabolites with 81% prediction accuracy. The strategy presented here contributes significantly to understand the pathology of UA. The paper presented a novel research avenue for investigating biological basis of syndrome. Logistic regression is a type of multivariable analysis used with increasing frequency in the health sciences because of its ability to model dichotomous outcomes. Proper use of this powerful and sophisticated modeling technique requires considerable care both in the specification of the form of the model and in the calculation and interpretation of the model‟s coefficients. The criteria considered in this article can affect the regression coefficients, in different ways and at different stages of the model-building process. Although many parts of the process have been effectively automated, the authority of the final model depends on the attempts by investigators to rule out sources of bias or inaccuracy toward which each of the criteria contributes.

Acknowledgment

We would like to thank our honorable principal Dr. Hari Vasudevan of D.J.Sanghvi College of Engineering and Head of Department of Computer Engineering, Dr.Narendra Shekhokar for giving us the facilities and providing us with a propitious environment for working in college. We would also like to thank S.V.K.M. for encouraging us in such co-curricular activities.

References

Vollmer RT (1996) Multivariate statistical analysis. Part I, The logistic model. Am J Clin Pathol;105:115–26.

Lemeshow S, Hosmer DW. (1998) Logistic regression. In: Armitage P, ColtonT, Eds. New York: J. Wiley,.p. 2316–27.

Glantz SA, Slinker BK. (1990) Primer of applied regression variance. New York: McGraw-Hill, Inc..

J. X Chen, G. C Xi, J. Chen, Y.S. Zhen, Y. W Xing, J. Wang, W. Wang. (2007) An unsupervised pattern (syndrome in traditional Chinese medicine) discovery algorithm based on association delineated by revised mutual information in chronic renal failure data. Journal of biological systems,;15: 435-51.

S. Z Guo, J. X Chen, H.H Zhao, W. Wang, J. Q Yi, L. Liu, Q. G Qi, R. Q Liu, Q. Qiu, Y ,H.H Zhao (2009), Acta Chimica Sinica.;67:167-73.

H.H Zhao, N. Hou, W. Wang. (2009) Difference Expressed Protein Study on Unstable Angina Blood-stasis Syndrome By Fluorescent Labeling Method. Spectroscopy and Spectral Analysis.29:1647-50

H.H Zhao, J. X Chen, N. Hou, P. Zhang, Y. Wang, J. Han, Q. Hou, Q. G. Qi, W. Wang. (2010) Discovery of Diagnosis Pattern of Coronary Heart Disease with Qi Deficiency Syndrome by T Test based Adaboost Algorithm. Evidence-based Compl. and Alt. Medicine;7:101-18.