Journal of Computer Science & Systems Biology - Open Access
Research Article JCSB/Vol.2 January-February 2009

Performance Comparative in Classification Algorithms Using Real Datasets

Hanuman Thota[1,2], Raghava Naidu Miriyala[1,2], Siva Prasad Akula[2,5], K.Mrithyunjaya Rao[3], Chandra Sekhar Vellanki[1], Allam Appa Rao[4], Srinubabu Gedela[4,5]*

[1]D.M.S S.V.H. College of Engineering, Department of Computer Science, Machilipatnam-521002, India.
[2]Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur-522510, India
[3]Vaagdevi College of Engineering, Warangal-506005, India.
[4]International Center for Bioinformatics, Department of Computer Science and Systems Engineering , Andhra University, Visakhapatnam-530003, India
[5]Institute of Glycoproteomics & Systems Biology, Tarnaka, Hyderbad-500017, India

*Corresponding author: Srinubabu Gedela, Institute of Glycoproteomics & Systems Biology, Tarnaka, Hyderbad-500017, India, Tel: +91-40-27006539; Fax: +91-40-40131662; E-mail: srinubabuau6@gmail.com

Received December 13, 2008; Accepted February 08, 2009; Published February 23, 2009

Citation: Hanuman T, Raghava NM, Siva PA, Mrithyunjaya RK , Chandra SV, et al. (2009) Performance Comparative in Classification Algorithms Using Real Datasets. J Comput Sci Syst Biol 2: 097-100. doi:10.4172/jcsb.1000021

Abstract

Classification is one of the most common data mining tasks, used frequently for data categorization and analysis in the industry and research. In real-world data mining sometimes it mainly deals with noisy information sources, because of data collection inaccuracy, device limitations, data transmission and discretization errors, or man-made perturbations frequently result in imprecise or vague data which is called as noisy data. This noisy data may decrease performance of any classification algorithms.. This paper deals with the performance of different classification algorithms and the impact of feature selection algorithm on Logistic Regression Classifier., How it controls False Discovery Rate (FDR) and thus improves the efficiency of Logistic Regression classifier.

Keywords: Classification; Data mining; Logistic regression classifier; Feature selection algorithm and false discovery Rate

Introduction

Data mining is the process of finding hidden patterns from data. It has wide range of applications such as, predicting stock prices, identifying suspected terrorists and scientific discovery like analysis of DNA microarray etc, as Shelke, et al., 2007 said researchers can now routinely investigate the biological molecular state of a cell measuring the simultaneous expression of tens of thousands of genes using DNA microarrays, Datamining can be used in the classification of proteins by basing on its primary structures (sequences) is presented. It contains four steps which include textmining, feature selection, datamining and classification. The sequences of protein are collected in a file (Mhamdi et al., 2004). In real-world data mining sometimes it mainly deals with noisy information sources, because of data collection inaccuracy, device limitations, data transmission and discretization errors, or man-

made perturbations frequently result in imprecise or vague data. Two common practices are to adopt either data cleaning approaches to enhance the data consistency or simply take noisy data as quality sources and feed them into the data mining algorithms. Either way may substantially sacrifice the mining performance(Wu and Zhu, 2008). noisy information plays critical role when making critical decision especially when using mining techniques in medical domain. The widespread availability of new computational methods and tools for data analysis and predictive modeling requires medical informatics researchers and practitioners to systematically select the most appropriate strategy to cope with clinical prediction problems, the collection of methods known as 'data mining' to deal with the analysis of medical data and construction of prediction models (Bellazzi and Zupan, 2008). Medical data mining is one of key issues to get useful clinical knowledge from medical databases. However, users often face difficulties during such medical data mining process for data preprocessing method selection/ construction, mining algorithm selection, and post-processing to refine the data mining process (Abe et al., 2007). Temporaldata miningis concerned with data mining of large sequential data sets ,there are application areas that need knowledge from temporal data such as sequential patterns, although there are many studies for temporal data mining, they do not deal with discovering knowledge from temporal interval data such as patient histories(Lee et al., 2009).

A commercial Web page typically contains many information blocks. Apart from the main content blocks, it usually has such blocks as navigation panels, copyright and privacy notices, and advertisements, these blocks are called noisy blocks. The information contained in these noisy blocks can seriously harm Web data mining (Tripathy and Singh, 2004). Noisy data is inherent in the field of data mining. If prior knowledge of such data was available, it would be a simple process to remove or account for noise and improve model robustness. Other wise its big disaster Unfortunately, in the majority of learning situations, the presence of underlying noise is suspected but difficult to detect (Liu et al., 2005).

The performance of a classification algorithm in data mining is greatly affected by the quality of data source. Irrelevant and redundant features of data not only increase the cost of mining process, but also degrade the quality of the result in some cases (Wu et al., 2006). Each algorithm has its own advantages and disadvantages , comparative study of Dr, Shuqing Huang is considered , as per his study and results by using UCI real data sets are used in the experiments.

Five groups of UCI real data sets are used. They are

i) Crab species data with 200 instances,5 attributes ,2 classes.

ii) Diabetes data with 768 instances, 8 attributes, 2 classes.

iii) Housing database with 506 instances, 12continuous attributes, 5classes .

iv) Iris plant database with 3 classes,4numeric attributes,150 instances.

v) Spambase database with 4601 instances,57 attributes and 2 classes.

These data sets are used from http://archive.ics.uci.edu/ml/index.htm to algorithms Decision tree, Support Vector Machine and Logistic Regression classifier.

Table 1 shows performance on different datasets available from the above mentioned repository.

As per the analysis support vector machine has more parameters than logistics regression and decision tree classifier, As shown in the experiment results, support vector

Real datasets	Decision tree	SVM	LRC
Crabs- species	87%	96%	82%
Diabetes	74%	88%	76%
Iris	96%	95%	87%
Housing-data	74%	93%	94%
Spambase-data	93%	91%	90%

Table 1: The interpretation of Classification result between Decision tree, Support Vector Machine and Logistic Regression classifier.

machine has the highest classification precision most of the time .However support vector machine is very time consuming because of more parameters ,demands more computation time. Compared to support vector machine, logistic regression is computationally efficient. Its results usually has static meaning. However it does not perform well when data set exhibits explicit data structures.

Feature Selection Based by Controlling False Discovery Rate

Contemporary biological technologies produce extremely high-dimensional data sets from which to design classifiers, with 20,000 or more potential features being common place. In addition, sample sizes tend to be small here, feature selection is an inevitable part of classifier design (Hua et al., 2009). The objective of feature selection is to get a feature subset that has the best performance (Cho, 2009) goal of feature selection is to select a subset of attributes from possible high dimensional feature space to a low dimensional spaces which is better suited for retrieval or learning purposes.

Bio-chip data that consists of high-dimensional attributes have more attributes than specimens. Thus, it is difficult to obtain covariance matrix from tens thousands of genes within a number of samples. Feature selection and extraction is critical to remove noisy features and reduce the dimensionality in microarray analysis(Lin and Chien, 2009). Research efforts have reported with increasing confirmation that the support vector machines (SVM) have greater accurate diagnosis ability (Akay, 2009). The assessment of risk factor of default on credit is important for financial institutions. Logistic regression technique traditionally used in credit scoring for determining likelihood to default based on consumer application and credit reference agency data. But support vector machines against these traditional methods on a large credit card database. Can be used as the basis of a feature selection method to discover those features that are most significant in determining risk of default.(Bellotti and Crook, 2009).

By applying feature selection on UCI real data sets were used in the experiments, five groups UCI real data sets were used. They are crab species data, diabetes data housing database , IRIS plant database, spam database. Classification results after applying false discovery rate (FDR), results taken from the observation of Dr, shuqing huang.

	Decision tree	SVM	LRC	LRC+FDR
Crabs- species	87%	96%	82%	85%
Diabetes	74%	88%	76%	77%
Iris	96%	95%	87%	97%
Housing-data	74%	93%	94%	95%
Spambase-data	93%	91%	90%	91%

Table 2: The interpretation of the result between Decision tree, Support Vector Machine and Logistic Regression classifier with feature selection.

The False Discovery Rate (FDR) of a set of predictions is the ***expected*** percent of false predictions in the set of predictions. For example if the algorithm returns 100 genes with a false discovery rate of 3 then we should expect 70 of them to be correct .Different features play different roles in classifying datasets. Unwanted features will result in error information during classification which will reduce classification precision. Feature selection can remove these distractions to improve classification performance. As shown in the experimental results, after feature selection using the proposed algorithm to control false discovery rate, the classification performance of logistic regression classifier was improved.

Conclusion

The experimental results show that when feature selection applied on Logistic Regression classifier controls False Discovery Rate and thus improves efficiency of Logistic Regression classifier , FDR controls the expected proportion of incorrectly rejected null hypotheses (errors) . Even though it cant be eradicated completely but can improve the efficiency of the regression classifier further.

Acknowledgement

We are extremely thankful to Dr Shqing Huang ,Tulane University, U.S.A and Dr.Raghu B. Korrapati, Walden University, U.S.A for

their literature. Dr.Rama kishore, M.Nancharaiah and V.Rakesh for their moral support.

References

1. Abe H, Yokoi H, Ohsaki M, Yamaguchi T (2007) Developing an integrated time-series data mining environment for medical data mining. Proceedings - IEEE International Conference on Data Mining, ICDM 4476657: pp127-132. » CrossRef » Google Scholar

2. Akay MF (2009) Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 36: pp3240-3247. » CrossRef » Google Scholar

3. Bellazzi R, Zupan B, (2008) Predictive data mining in clinical medicine: Current issues and guidelines. Int J Med Inform 77: pp81-97. » CrossRef » Pubmed » Google Scholar

4. Bellotti T, Crook J (2009) Support vector machines for credit scoring and discovery of significant features. Expert Syst Appl 36: pp3302-308. » CrossRef » Google Scholar

5. Cho HW (2009) A data mining-based subset selection for enhanced discrimination using iterative elimination of redundancy. Expert Syst Appl 36: pp1355-1361. » CrossRef » Google Scholar

6. Hua J, Tembe WD, Dougherty ER (2009) Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognit 42: pp409-424.

7. Lee YJ, Lee JW, Chai DJ, Hwang BH, Ryu KH (2009) Mining temporal interval relational rules from temporal data. Journal of Systems and Software 82: pp155-167.

8. Lin KS, Chien CF (2009) Cluster analysis of genome-wide expression data for feature extraction. Expert Syst Appl 36: pp3327-3335.

9. Liu XD, Shi CY, Gu XD (2005) A boosting method to detect noisy data. 2005 International Conference on Machine Learning and Cybernetics, ICMLC 2005, pp2015-2020. » CrossRef » Google Scholar

10. Mhamdi F, Elloumi M, Rakotomalala R, (2004) Textmining, feature selection and datamining for proteins classification. Proceedings - 2004 International Conference on Information and Communication Technologies: From Theory to Applications, ICTTA 2004 pp457-458. » CrossRef » Google Scholar

11. Shelke RR, Deshmukh VM (2007) Computational analysis of DNA microarray data using datamining. Biosci Biotechnol Res Asia 4: pp321-324. » Google Scholar

12. Tripathy AK, Singh AK (2004) An efficient method of eliminating noisy information in web pages for data mining. Proceedings - The Fourth International Conference on Computer and Information Technology (CIT 2004) pp978-985.

13. Wu W, Gao Q, Wang M (2006) An efficient feature selection method for classification data mining. WSEAS Transactions on Information Science and Applications 3: pp 2034-2040.

14. Wu X, Zhu X (2008) Mining with noise knowledge: Error-aware data mining. IEEE Trans Syst Man Cyybern A Syst Hum 38: pp917-932.