ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Booster in High Dimensional Data Classification with Q-Statistic

Mallampati Visweswararao
M.Tech
Department of CSE,
MJR College of Engineering & Technology,
Piler, Andhra Pradesh.

Varaprasad Gajjala
Assistant Professor
Department of CSE,
MJR College of Engineering & Technology,
Piler, Andhra Pradesh.

Karamala Suresh
HoD
Department of CSE,
MJR College of Engineering & Technology,
Piler, Andhra Pradesh.

Abstract:
Classification problems is a great challenge in high dimensional data with number of observations are becoming more common in microarray data. Last two decades, lots of efficient classification problems and feature selection (FS) algorithms have been implemented for higher prediction accuracies. Feature selection algorithm results in high prediction accuracy for classification but the result is not stable when training set differs, eminently in high dimensional data. The large amount of text information on the Internet web pages affects the clustering analysis. The text clustering is a favorable analysis technique used for partitioning a massive amount of information into clusters. Hence, the major problem that affects the text clustering technique is the presence uninformative and sparse features in text documents. A broad class of boosting algorithms can be interpreted as performing coordinate-wise gradient descent to minimize some potential function of the margins of a data set. This paper proposes a new evaluation measure Q-statistic that incorporates the stability of the selected feature subset in addition to the prediction accuracy. Here different micro array real data sets is used to show that booster not only boost the prediction accuracy but also boost the Q –statistic. Micro array data is a collection of similar data. Hence dealing with high dimensional data is very difficult for classification Feature Selection with boosting technique is applied for improving accuracy.

KEYWORDS: ***high dimensional data classification; feature selection; stability; Q-statistic; Booster;***

Introduction

The presence of high dimensional data is becoming more common in many practical applications such as data mining, machine learning and micro array gene expression data analysis. Typical publicly available micro array data has tens of thousands of features with small sample size and the size of the features considered in microarray data analysis is growing. Recently, after the increasing amount of digital text on the Internet web pages, the text clustering (TC) has become a hard technique used to clustering a massive amount of documents into a subset of clusters. It is used in the area of the text mining, pattern recognition and others. Vector Space Model (VSM) is a common model used in the text mining area to represents document components. Hence, each document is represented as a vector of terms weight, each term weight value is represented as a one dimension space. Usually, text documents contain informative and uninformative features, where an uninformative is as irrelevant, redundant, and uniform distribute features. Unsupervised feature section (FS) is an important task used to find a new subset of informative features to improve the TC algorithm.

Methods used in the problems of statistical variable selection such as forward selection, backward elimination and their combination can be used for FS problems. Most of the successful FS algorithms in high dimensional problems have utilized forward selection method but not considered backward elimination method since it is impractical to implement backward elimination process with huge number of features. The statistical classification of the data with large number of features and small sample size (under sampled problem) presents an intrinsic challenge. A striking result has been found that the simple and popular Fisher linear discriminate analysis can be as poor as random guessing as the number of features gets larger. Most of the features of high dimensional microarray data are irrelevant to the target feature and the proportion of relevant features or the percentage of up-regulated or down-regulated genes compared with appropriate normal tissues is only 2% _ 5%. Finding relevant features simplifies learning process and increases prediction accuracy. The finding, however, should be relatively robust to the variations in training data, especially in biomedical study, since domain experts will invest considerable time and efforts on this small set of selected features. Hence, the proposed selection should provide them not only with the high predictive potential but also with the high stability in the selection.

New Proposal for Feature Selection

This paper proposes Q-statistic to evaluate the performance of an FS algorithm with a classifier. This is a hybrid measure of the prediction accuracy of the classifier and the stability of the selected features. Then the paper proposes Booster on the selection of feature subset from a given FS algorithm. The basic idea of Booster is to obtain several data sets from original data set by re-sampling on sample space. Then FS algorithm is applied to these re-sampled data sets to obtain different feature subsets. The union of these selected subsets will be the feature subset obtained by the Booster of FS algorithm. Experiments were conducted using spam email. The authors found that the proposed genetic algorithm for FS is improved the performance of the text. The FS technique is a type of optimization problem, which is used to obtain a new subset of features. Cat swarm optimization (CSO) algorithm has been proposed to improve optimization problems. However, CSO is restricted to long execution times. The authors modify it to improve the FS technique in the text classification. Experiment Results showed that the proposed modified CSO overcomes tradition al version and got more ace uprate results in FS technique.

Booster

Booster is simply a union of feature subsets obtained by a re-sampling technique. The re-sampling is done on the sample space. Three FS algorithms considered in this paper are minimal-redundancy-maximal-relevance, Fast Correlation-Based Filter, and Fast clustering-based feature Selection algorithm. All three methods work on discretized data. For mRMR, the size of the selection m is fixed to 50 after extensive experimentations. Smaller size gives lower accuracies and lower values of Q-statistic while the larger selection size, say 100, gives not much improvement over 50. The background of our choice of the three methods is that FAST is the most recent one we found in the literature and the other two methods are well known for their efficiencies. FCBF and mRMR explicitly include the codes to remove redundant features. Although FAST does not explicitly include the codes for removing redundant features, they should be eliminated implicitly since the algorithm is based on minimum spanning tree. Our extensive experiments supports that the above three FS algorithms are at least as efficient as other algorithms including CFS.

Existing System

Methods used in the problems of statistical variable selection such as forward selection, backward elimination and their combination can be used for FS problems. Most of the successful FS algorithms in high dimensional problems have utilized forward selection method but not considered backward elimination method since it is impractical to implement backward elimination process with huge number of features. A serious intrinsic problem with forward selection is, however, a flip in the decision of the initial feature may lead to a completely different feature subset and hence the stability of the selected feature set will be very low although the selection may yield very high accuracy.

This is known as the stability problem in FS. The research in this area is relatively a new field and devising an efficient method to obtain a more stable feature subset with high accuracy is a challenging area of research. Disadvantages of existing system is several studies based on re-sampling technique have been done to generate different data sets for classification problem, and some of the studies utilize re-sampling on the feature space.

Proposed System

This paper proposes Q-statistic to evaluate the performance of an FS algorithm with a classifier. This is a hybrid measure of the prediction accuracy of the classifier and the stability of the selected features. Then the paper proposes Booster on the selection of feature subset from a given FS algorithm. The basic idea of Booster is to obtain several data sets from original data set by re-sampling on sample space. Then FS algorithm is applied to each of these re-sampled data sets to obtain different feature subsets. The union of these selected subsets will be the feature subset obtained by the Booster of FS algorithm. Empirical studies show that the Booster of an algorithm boosts not only the value of Q-statistic but also the prediction accuracy of the classifier applied. There are some advantages in proposed system: 1) the prediction accuracy of classification without consideration on the stability of the selected feature subset. 2) The MI estimation with numerical data involves density estimation of high dimensional data.

Efficiency of Booster

There are two concepts in Booster to reflect the two domains. The first is the shape, Booster’s equivalent of a traditional array[6] a finite set of elements of a certain data-type, accessible through indices. Unlike arrays, shapes need not necessarily be rectangular for convenience we will, for the moment, assume that they are. Shapes serve, from the algorithm designer’s point of view, as the basic placeholders for the algorithm’s data: input-, output-, and intermediate values are stored within shapes. As we will see later on, this does not necessarily mean that they are represented in memory that way, but the algorithm designer is allowed to think so. It presents the effect of s-Booster on accuracy and Q-statistic against the original.

Booster Boost Accuracy

Boosting is a technique for generating and combining multiple classifiers to improve predictive accuracy. It is a type of machine learning meta-algorithm for reducing bias in supervised learning and can be viewed as minimization of a convex loss function over a convex set of functions. At issue is whether a set of weak learners can create a single strong learner A weak learner is defined to be a classifier which is only slightly correlated with the true classification and a strong learner is a classifier that is arbitrarily well-correlated with the true classification. Learning algorithms that turn a set of weak learners into a single strong learner is known as boosting.

Booster Boost Q-Statistic

Q-static search algorithm generates random memory solutions and pursuing to improve the harmony memory to obtain optimal solution an optimal subset of informative features. Each musician unique term is a dimension of the search space. The solutions are evaluated by the fitness function as it is used to obtain an optimal harmony global optimal solution. Harmony search algorithm performs the fitness function is a type of evaluation criteria used to evaluate solutions. At each iteration the fitness function is calculated for each HS solution. Finally, the solution, which has a higher fitness value, is the optimal solution. We used mean absolute difference as fitness function in HS algorithm for FS technique using the weight scheme as objective function for each position.

System Architecture

A well-planned data classification system makes essential data easy to find and retrieve. This can be of particular importance for and written procedures and guidelines for data classification should define what categories and criteria the organization will use to classify data and specify the roles and responsibilities of

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

employees within the organization regarding. Once a data-classification scheme has been created, security standards that specify appropriate handling practices for each category and storage standards that define the requirements should be addressed. To be effective, a classification scheme should be simple enough that all employees can execute it properly. Here is an example of what a data classification.

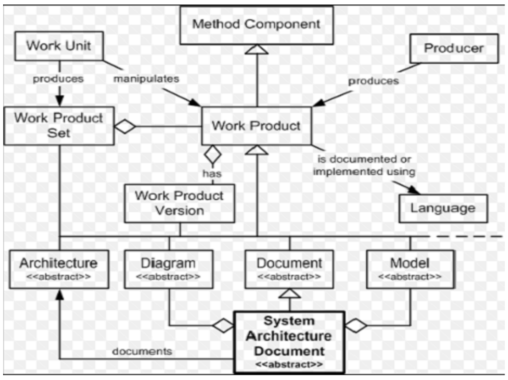

Fig 1. System Architecture

Experiment Description

All of these data sets have their own properties like the domain of the data set, the kind of attributes it contains, and tree size after training. We tested each data set with four different classification tree algorithms: J48, REP Tree, Random Tree and Logistical Model Trees. For each algorithm both the test options percentage split and cross-validation were used. With percentage split, the data set is divided in a training part and a test part. For the training set 66% of the instances in the data set is used and for the test set the remaining part. Cross-validation is especially used when the amount of data is limited. Instead of reserving a part for testing, cross-validation.

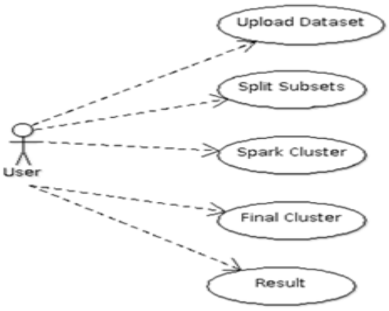

Fig 2 : Use case module

Simulation Results

In this boosting it will show the exact difference between accurate and non accurate boosting. Early stopping cannot save a boosting algorithm it is possible that the global optimum analyzed in the preceding section can be reached after the first iteration. Since depends only on the inner product between and the normalized example vectors, it follows that rotating the set S around the origin by any fixed angle induces a corresponding rotation of the function and in particular of its minima. Note that we have used here the fact that every example point in S lies within the unit disc; this ensures that for any rotation of S each weak hypothesis xi will always give outputs in as required. Consequently a suitable rotation to will result in the corresponding rotated function having a global minimum at a vector which lies on one of the two coordinates.

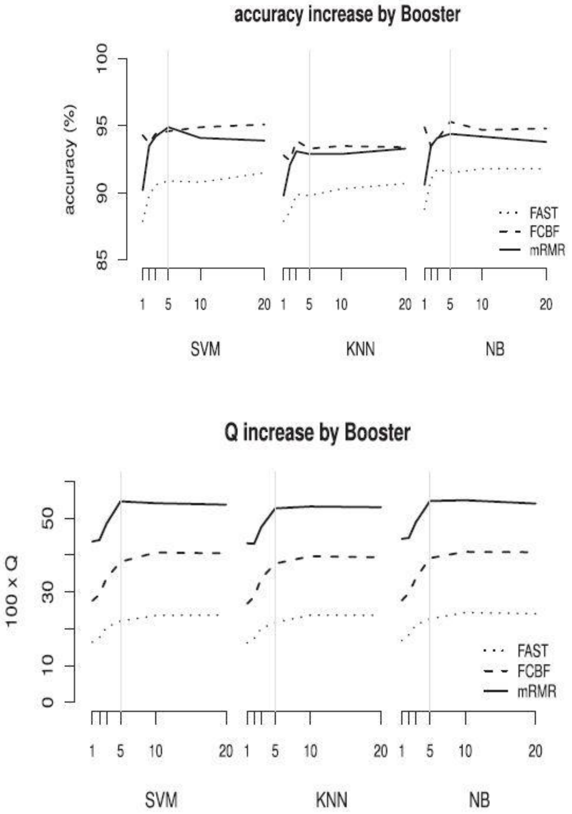

Fig. 3. Accuracy and Q-statistic of s Boosterb for b ¼ 1; 2; 3; 5; 10; and 20 (x-axis). Each value is the average over the 14 data sets. s-Booster1 is s. The grey vertical line is for b ¼ 5.

Conclusion

This proposed a measure Q-statistic that evaluates the performance of an FS algorithm. Q-statistic accounts both for the stability of selected feature subset and the prediction accuracy. The paper proposed Booster to boost the performance of an existing FS algorithm. Experimentation with synthetic data and microarray data sets has shown that the suggested Booster improves the prediction accuracy and the Q-statistic of the three well-known FS algorithms: FAST, FCBF, and mRMR. Also we have noted that the classification methods applied to Booster do not have much impact on prediction accuracy and Q-statistic. Our results show, for the four classifications tree algorithms we used, that using cost-complexity pruning has a better performance than reduced-error pruning. But as we said in the results section, this could also be caused by the classification algorithm itself. To really see the difference in performance in pruning methods another experiment can be performed for further/future research. Tests could be run with algorithms by enabling and disabling the pruning option and using more different pruning methods. This can be done for various classification tree algorithms which use pruning. Then the increase of performance by enabling pruning could be compared between those classification tree algorithms.

References

1. A. Criminisi, P. Perez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting.”, IEEE Transactions on Image Processing, vol. 13, no.9, pp. 1200–1212, 2004.

2. Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, Stanley Osher, “Simultaneous Structure and Texture Image Inpainting”, IEEE Transactions On Image Processing, vol. 12, No. 8, 2003.

3. Yassin M. Y. Hasan and Lina J. Karam, “Morphological Text Extraction from Images”,IEEE Transactions On Image Processing, vol. 9, No. 11, 2000.

4. Eftychios A. Pnevmatikakis, Petros Maragos “An Inpainting System For Automatic Image Structure-Texture Restoration With Text Removal”, IEEE trans. 978-1-4244-1764, 2008.

5. S.Bhuvaneswari, T.S.Subashini, “Automatic Detection and Inpainting of Text Images”, International Journal of Computer Applications (0975 – 8887) Volume 61– No.7, 2013.

6. G. H. John, R. Kohavi, and K. Pfleger,” Irrelevant features and the subset selection problem,” in Proc. 11th Int. Conf. Mach. Learn., vol. 94, pp. 121–129, 1994.

7. G. H. John and P. Langley, “Estimating continuous distributions in Bayesian classifiers,” in Proc. 11th Conf. Uncertainty Artif. Intell., pp. 338–345, 1995.

8. R. V. Jorge and A. E. Pablo, “A review of feature selection methods based on mutual information,” Neural Comput. Appl., vol. 24, no. 1, pp. 175–186, 2014.

9. A. Kalousis, J. Prados, and M. Hilario, “Stability of feature selection algorithms: A study on high-dimensional spaces,” Knowl. Inf. Syst., vol. 12, no. 1, pp. 95–116, 2007.

10. I. Kojadinovic, “Relevance measures for subset variable selection in regression problems based on k-additive mutual information,” Comput. Statist. Data Anal., vol. 49, no. 4, pp. 1205–1227, 2005.

11. D. Koller and M. Sahami, “Toward optimal feature selection,” in Proc. 13th Int. Conf. Mach. Learn., pp. 284–292, 1996.

12. L. I. Kuncheva, “A stability index for feature selection,” in Proc. Artif. Intell. Appl., pp. 421–427, 2007.

13. H. Liu, J. Li, and L.Wong, “A comparative study on feature selection and classification methods using gene

expression profiles and proteomic patterns," Genome Informatics Series, vol. 13, pp. 51–60, 2002.

14. H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L.Wasserman, "Forest density estimation," The J. Mach. Learn. Res., vol. 12, pp. 907–951, 2011.

15. R. S. Marko, and I. Kononenko, "Theoretical and empirical analysis of, vol. 53, no. 1–2, pp. 23–69, 2003.

Author's Profile:

Mallampati Visweswararao received M.Tech Degree from MJR College of Engineering & Technology, Department of CSE, Piler, Andhra Pradesh.
Email : rao1206@gmail.com

Varaprasad Gajjala Working as a Assistant Professor in MJR College Of Engineering And Technology, Piler, A.P. and he obtained Graduation Degree in Computer Science and Engineering from JNTUA university in 2012,and obtained his Master's degree in 2014 specialized in Artificial Intelligence under Department of Computer Science & Engineering JNTUA college of Engineering, Anantapuramu, India. He made robot based on Artificial Intelligence while pursuing Master's. His research interest includes Web Designing & Graphic Designing, Information Security, Image Processing, Natural Language Processing, Robotics, and Artificial Intelligence.
Email: gajjala.varaprasad@gmail.com

Karamala Suresh Working as a Head of the Department In MJR College Of Engineering And Technology, Department Of Cse,Piler ,Chittoor dist. He is having 14 years of teaching experience in engineering colleges, he received B.Tech(CSE) from JNTU Hyderabad in 2002,Received M.E(CSE) from satyabama university Chennai in 2006,and Received M.Tech(CSE) in 2015 from JNTU Anantapuramu, he published several research papers in various national and international journals. He is interested in Computer Networks.