

Improving Efficiency Using Q-Statistics and FS Algorithm in High Dimensional Datasets

¹P.V.N Rajeswari, ²Nataraja Priyanka

¹Assoc. Prof in Dept. of CSE, PBR Visvodaya Institute of Technology & Science, Kavali, AP, India.

²M.Tech, Dept. of CSE, PBR Visvodaya Institute of Technology & Science, Kavali, AP, India.

Abstract – Classification problems in high dimensional data with small number of observations are becoming more common especially in microarray data. The increasing amount of text information on the Internet web pages affects the clustering analysis [1]. 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 of 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 [1]. This paper proposes a new evaluation measure Q-statistic that incorporates the stability of the selected feature subset in addition to the prediction accuracy. Then we propose the Booster of an FS algorithm that boosts the value of the Q statistic of the algorithm applied.

Keywords: High dimensional data classification; feature selection; stability; Q-statistic; Booster;

I. INTRODUCTION

The presence of high dimensional data is becoming more common in much practical application such as data mining, machine learning and micro array gene expression data analysis. Typical publicly available microarray data has tens of thousands of features with small sample size and the size of the features considered in microarray data analysis is growing[1][2]. 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 [3]. 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.

II. BACKGROUND WORK

In the year of 2014, the authors Y. Wang, L. Chen, and J.-P. Mei. Revealed a paper titled "Incremental fuzzy clustering with multiple medoids for large data" and describe into the paper such as a critical strategy of information investigation, grouping assumes an essential part in finding the fundamental example structure installed in unlabeled information. Grouping calculations that need to store every one of the information into the memory for examination get to be distinctly infeasible when the dataset is too vast to be put away. To handle such extensive information, incremental bunching methodologies are proposed.

The point by point issue definition, overhauling rules determination, and the top to bottom investigation of the proposed IMMFC are given. Trial examines on a few huge datasets that incorporate genuine malware datasets have been led. IMMFC outflanks existing incremental fluffy bunching approaches as far as grouping exactness and power to the request of information. These

outcomes show the colossal capability of IMMFC for huge information examination.

Clustering is proposed, for automatically exploring potential clusters in dataset. This uses supervised classification approach to achieve the unsupervised cluster analysis. Fusion of clustering and fuzzy set theory is nothing but fuzzy clustering, which is appropriate to handle problems with imprecise boundaries of clusters. A fuzzy rule-based classification system is a special case of fuzzy modeling, in which the output of system is crisp and discrete. Fuzzy modeling provides high interpretability and allows working with imprecise data. To explore the clusters in the data patterns, FRBC appends some randomly generated auxiliary patterns to the problem space. It then uses the main data as one class and the auxiliary data as another class to enumerate the unsupervised clustering problem as a supervised classification one.

III. A 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 resampling on sample space. Then FS algorithm is applied to these resampled data sets to obtain [4], [5] 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 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 traditional version and got more accuracy results in FS technique.

IV. BOOSTER

Booster is simply a union of feature subsets obtained by a resampling technique. The resampling 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.[6] 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 support that the above three FS algorithms are at least as efficient as other algorithms including CFS.

V. 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 originals. Classifier used here is NB.

a) *Booster Boost S 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.

b) *Booster Boosts 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.

VI. 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 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.

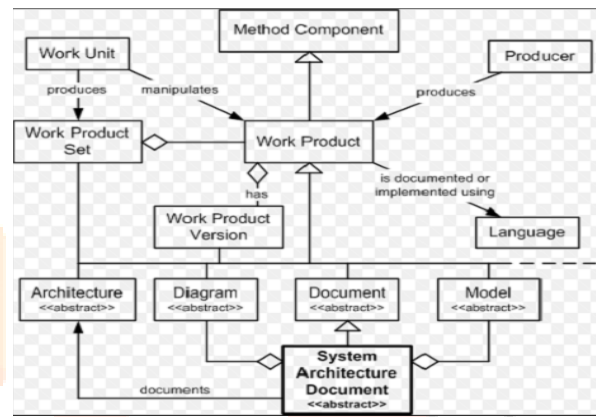


Fig 1. System Design

VII. 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] Aria Pezeshk and Richard L. Tutwiler, "Automatic Feature Extraction and Text Recognition from Scanned Topographic Maps", IEEE Transactions on geosciences and remote sensing, VOL. 49, NO. 12, 2011.
- [7] Xiaoqing Liu and Jagath Samarabandu, "Multiscale Edge-Based Text Extraction From Complex Images", IEEE Trans., 1424403677, 2006
- [8] Nobuo Ezaki, Marius Bulacu Lambert, Schomaker, "Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons", Proc. of 17th Int. Conf. on Pattern Recognition (ICPR), IEEE Computer Society, pp. 683-686, vol. II, 2004.
- [9] G. Brown, A. Pocock, M. J. Zhao, and M. Lujan, "Conditional likelihood maximization: A unifying framework for information theoretic feature selection," J. Mach. Learn. Res., vol. 13, no. 1, pp. 27–66, 2012.
- [10] C. Kamath, Scientific data mining: a practical perspective, Siam, 2009
- [11] B. C. Christensen, E. A. Houseman, C. J. Marsit, S. Zheng, M. R. Wrensch, H. H. Nelson, M. R. Karagas, J. F. Padbury, R. Bueno, D. J. Sugarbaker, R.-F. Yeh, J. K. Wiencke, and K. T. Kelsey, "Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context," PLOS Genetics, vol. 5, no. 8, e1000602, 2009.
- [12] C. Corinna and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.
- [13] T. M. Cover and J. A. Thomas, Elements of Information Theory (Series in Telecommunications

and Signal Processing), 2nd ed. Hoboken, NJ, USA: Wiley, 2002.

- [14] D. Dembele, "A flexible microarray data simulation model," *Microarrays*, vol. 2, no. 2, pp. 115–130, 2013.
- [15] D. Derroncourt, B. Hanczar, and J. D. Zucker, "Analysis of feature selection stability on high dimension and small sample data," *Comput. Statist. Data Anal.*, vol. 71, pp. 681–693, 2014.

AUTHORS

P.V.N Rajeswari has received her B.Tech in Computer Science & Engineering and M.Tech degree in Computer science & Engineering from Andhra University in 2004 and Allahabad University in 2006 respectively. She is also pursuing PhD in Andhra University. She is dedicated to teaching field from the last 13 years. She has guided 18 P.G and 42.UG students. Her research areas included Data Mining., Machine Learning and Big Data. At present she is working as Associate Professor in Visvodaya Engineering College, Kavali, Andhra Pradesh, India.

Nataraja Priyanka has received her B.Tech in Information Technology from Brahmas Institute of Engineering College, Nellore affiliated to JNTU, Anantapur in 2013 and pursuing M.Tech degree in Computer Science and Engineering in PBR Visvodaya Institute of Technology and Science, Kavali, A.P affiliated to JNTU, Anantapur in (2015-2017).

