© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Booster of an FS Algorithm on High Dimensional Data

N.Hima Bindu[1], T.Chakravarthi[2]

[1]M.Tech Scholar Department of CSE, NRI Institute of Technology Visadala(V&M),Guntur(Dt), Andhra Pradesh, India

[2]Assistant Professor Department of CSE, NRI Institute of Technology Visadala(V&M),Guntur(Dt), Andhra Pradesh, India

ABSTRACT

Classification issues in high dimensional knowledge with tiny variety of observations have become additional common particularly in microarray knowledge. The increasing quantity of text info on the net sites affects the agglomeration analysis[1]. The text agglomeration may be a favorable analysis technique used for partitioning a colossal quantity of knowledge into clusters. Hence, the most important downside that affects the text agglomeration technique is that the presence uninformative and distributed options 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, KDD, Preprocessing, Neural Networks, Decision trees.

I. INTRODUCTION

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, physician may make a decision based on the selected features whether a dangerous surgery is necessary for treatment or not. Feature selection in supervised learning has been well studied, where the main goal is to find a feature subset that produces higher classification accuracy. Recently, several researches (Dy and Brodley, 2000b, Devaney and Ram, 1997, Agarwal et al., 1998) have studied feature selection and clustering together with a single or unified criterion. For feature selection in unsupervised learning, learning algorithms are designed to find natural grouping of the examples in the feature space. Thus feature selection in unsupervised learning aims to find a good subset of features that forms high quality of clusters for a given number of clusters. However, the traditional approaches to feature selection with single evaluation criterion have shown limited capability in terms of knowledge discovery and decision support. This is because decision-makers should take into account multiple, conflicted objectives simultaneously. In particular no single criterion for unsupervised feature selection is best for every application (Dy and Brodley, 2000a) and only the decision maker can determine the relative weights of criteria for her application. In order to provide a clear picture of the (possibly nonlinear) tradeoffs among the various objectives, feature selection has been formulated as a multi-objective or Pareto optimization. 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 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. LITERATURE SURVEY

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 projected, for mechanically exploring potential clusters in dataset. This uses supervised classification approach to attain the unsupervised cluster analysis. Fusion of bunch and fuzzy pure mathematics is nothing however fuzzy bunch, that is suitable to handle issues with imprecise boundaries of clusters. A fuzzy rule-based organization may be a special case of fuzzy modeling, within which the output of system is crisp and distinct. Fuzzy modeling provides high interpretability and permits operating with imprecise knowledge. To explore the clusters within the knowledge patterns, FRBC appends some every which way generated auxiliary patterns to the matter house. It then uses the most knowledge in concert category and therefore the auxiliary knowledge as another category to enumerate the unsupervised bunch downside as a supervised classification one.

III. A NEW PROPOSAL FOR FEATURE SELECTION

This paper proposes Q-statistic to gauge the performance of AN FS rule with a classifier. this can be a hybrid live of the prediction accuracy of the classifier and therefore the stability of the chosen options. Then the paper proposes Booster on the choice of feature set from a given FS rule. The basic plan of Booster is to get many information sets from original information set by resembling on sample house. Then FS rule is applied to those resample information sets to obtain[4][5] totally different feature subsets. The union of those hand-picked sets are the feature subset obtained by the Booster of FS rule. Experiments were conducted victimization spam email. The authors found that the planned genetic rule for FS is improved the performance of the text. The FS technique could be a style of improvement downside, that is employed to get a replacement set of options. Cat swarm improvement (CSO) rule has been planned to enhance improvement issues. However, CSO is restricted to long execution times. The authors modify it to enhance the FS technique within the text classification. Experiment Results showed that the planned changed CSO overcomes tradition al version and got additional ace up rate leads to 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 discredited 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.

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.

A.BOOSTER BOOSTS 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.BOOSTERBOOSTS 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. 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 classification 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.

VII. REFERENCES

[1]. I.H. Witten, E. Frank and M.A. Hall, Data mining practical machine learning tools and techniques, Morgan Kaufmann publisher, Burlington 2011

[2]. J. Han and M. Kamber, Data mining concepts and techniques, Morgan Kaufmann, San Francisco 2006

[3]. T.J. Shan, H. Wei and Q. Yan, "Application of genetic algorithm in data mining", 1st Int Work Educ Technol Comput Sci, IEEE 2, 2009, pp. 353- 356

[4]. Z.Z. Shi, Knowledge discovery, Tsinghua University Press, Beijing, 2001

[5]. D. Pyle, Data preparation for data mining, 1st Vol., Morgan Kaufmann publisher, San Francisco, 1999

[6]. I. Guyon, N. Matic and V. Vapnik, "Discovering informative patterns and data cleaning", In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (ed) Advances in knowledge discovery and data mining, AAAI/MIT Press, California, 1996, pp. 181- 203

[7]. E. Simoudis, B. Livezey B and R. Kerber R , "Integrating inductive and deductive reasoning for data mining", In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (Eds.) Advances in knowledge discovery and data mining, AAAI/MIT Press, California, 1996, pp. 353-373

[8]. B. Pfahringer, "Supervised and unsupervised discretization of continuous features", Proc. 12th Int. Conf. Machine Learning, 1995, pp. 456-463.

[9]. J. Catlett, "On changing continuous attributes into ordered discrete attributes", In Y. Kodratoff (ed), Machine Learning—EWSL-91, Springer-Verlag, New York,1991, pp 164-178

[10]. W. Daelemans, V. Hoste, F.D. Meulder and B. Naudts, "Combined Optimization of Feature Selection and Algorithm Parameter Interaction in Machine Learning of Language", Proceedings of the 14th European Conference on Machine Learning (ECML-2003), Lecture Notes in Computer Science 2837, Springer-Verlag, Cavtat-Dubrovnik, Croatia, 2003, pp. 84-95

[11]. M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn and A.K. Jain, "Dimensionality Reduction

Using Genetic Algorithms", IEEE Transactions On Evolutionary Computation, Vol. 4, No. 2, 2000

[12]. Y. Saeys, I. Inza and P. Larranaga, "A review of feature selection techniques in bioinformatics", Bioinformatics-19, 2007, pp. 2507–17.

[13]. G.L. Pappa and A.A. Freitas, Automating the Design of Data Mining Algorithms. An Evolutionary Computation Approach, Natural Computing Series, Springer, 2010

[14]. A. Darwiche, Modeling and Reasoning with Bayesian Networks, Cambridge University Press, 2009

[15]. G.F. Cooper, P. Hennings-Yeomans, S. Visweswaran and M. Barmada, "An Efficient Bayesian Method for Predicting Clinical Outcomes from Genome-Wide Data", AMIA 2010 Symposium Proceedings, 2010, pp. 127-131

[16]. M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, "Building Decision Trees with Constraints", Data Mining and Knowledge Discovery, vol. 7, no. 2, 2003, pp. 187 – 214

[17]. T.M. Mitchell, Machine Learning, McGraw-Hill Companies, USA, 1997

[18]. Y. Singh Y, A.S. Chauhan, "Neural Networks in Data Mining", Journal of Theoretical and Applied Information Technology, 2005, pp. 37-42