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

Implementation of Clustering-Based Feature Subset Selection Algorithm for Large Data

K.Sireesha
M.Tech Student,
Department of CSE,
Sree Rama institute of Technology and Science,
Kuppenakuntla, Penuballi, Khammam,TS India.

S.Suresh
Assistant Professor,
Department of CSE,
Sree Rama institute of Technology and Science,
Kuppenakuntla, Penuballi, Khammam,TS India.

ABSTRACT:

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. Whilethe efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent, the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, FCBF, Relief, CFS, Consist, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the probability based Naive Bayes, the tree-based C4.5, the instance based IB1, and the rule-based RIPPER before and after feature selection. The results, on 35 publicly available real-world high dimensional image, microarray, and text data, demonstrate that the FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.

Index Terms:

Feature subset selection, filter method, feature clustering, graph-based clustering

INTRODUCTION:

WITH the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories . Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches.

The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning. algorithms, with good generality. Their computational complexity is low, but the accuracy of the learningalgorithms is not guaranteed . The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper.

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

They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper.

Existing System:

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features . Our proposed FAST algorithm falls into the second group. Traditionally feature subset selection research hasfocused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

Proposed System:

Recently, hierarchical clustering has been adopted in word selection in the context of text classification Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira et al.or on the distribution of class labels associated with each word by Baker and McCallum . As distributional clustering of words are agglomerative in nature, and result in suboptimal word clusters and high computational cost, Dhillon et al. proposed a newInformation theoretic divisive algorithm for word clustering and applied it to text classification.

Butterworth et al. proposed to cluster features using a special metric of Barthelemy-Montjardet distance, and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluationmeasure based on Barthelemy-Montjardet distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Further more, even compared with other feature selection methods, the obtained accuracy is lower.

FEATURE SUBSET SELECTION ALGORITHM

3.1 Framework and Definitions:

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines .Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant.

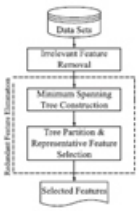

Fig. 1. Framework of the proposed feature subset selection algorithm.

and redundant features, and obtain a good feature subset. We achieve this through a new feature selection framework (shown in Fig. 1) which composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

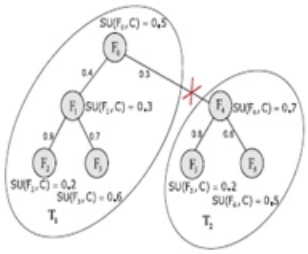

Fig. 2. Example of the clustering step.

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

Results and Analysis:

In this section, we present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy, and the Win/Draw/Loss record. For the purpose of exploring the statistical significance of the results, we performed a nonparametric Friedman test followed by Nemenyi post-hoc test as advised by Demsar and Garcia and Herrerato to statistically compare algorithms on multiple data sets. Thus, the Friedman and the Nemenyi test results are reported as well.

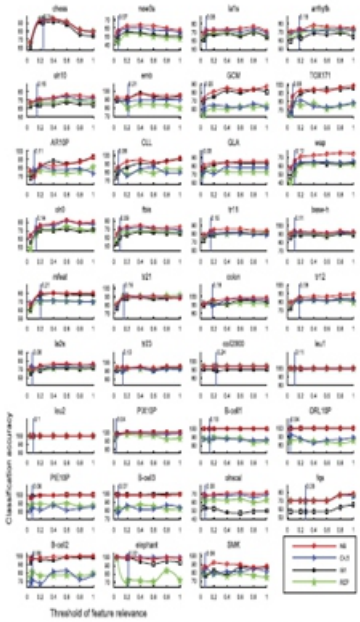

Fig. 9. Accuracies of the four classification algorithms with different θ values.

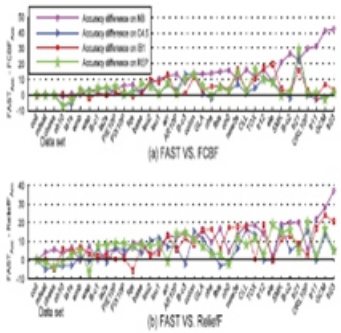

Fig. 10. Accuracy differences between FAST and the comparing algorithms.

CONCLUSION:

In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevantfeatures, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selectingrepresentative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. We have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, ReliefF, CFS, Consist, and FOCUS-SF on the 35 publicly available image, microarray,and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the bestproportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, C4.5, and RIPPER, and the second best classification accuracy for IB1. The Win/Draw/Loss records confirmed the conclusions We also found that FAST obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification accuracy of the four different types of classifiers, and CFS is a good alternative. At the same time, FCBF is a good alternative for image and text data. Moreover, Consist, and FOCUS-SF are alternatives for text data.

ACKNOWLEDGMENTS:

The authors would like to the editors and the anonymous reviewers for their insightful and helpful comments and suggestions, which resulted in substantial improvements to this work. This work is supported by the National Natural Science Foundation of China under grant 61070006.

REFERENCES:

[1] H. Almuallim and T.G. Dietterich, “Algorithms for IdentifyingRelevant Features,” Proc. Ninth Canadian Conf. Artificial Intelligence,pp. 38-45, 1992.

[2] H. Almuallim and T.G. Dietterich, “Learning Boolean Concepts inthe Presence of Many Irrelevant Features,” Artificial Intelligence,vol. 69, nos. 1/2, pp. 279-305, 1994.

[3] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, “A Feature SetMeasure Based on Relief,” Proc. Fifth Int’l Conf. Recent Advances inSoft Computing, pp. 104-109, 2004.

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

[4] L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.

[5] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[6] D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.

[7] J. Biesiada and W. Duch, "Features Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in Soft Computing, vol. 45, pp. 242-249, 2008.

[8] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584, 2005.

[9] C. Cardie, "Using Decision Trees to Improve Case-Based Learning," Proc. 10th Int'l Conf. Machine Learning, pp. 25-32, 1993.

[10] P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, "Mining of Attribute Interactions Using Information Theoretic Metrics," Proc. IEEE Int'l Conf. Data Mining Workshops, pp. 350-355, 2009.

[11] S. Chikhi and S. Benhammada, "ReliefMSS: A Variation on aFeature Ranking Relieff Algorithm," Int'l J. Business Intelligence andData Mining, vol. 4, nos. 3/4, pp. 375-390, 2009.

[12] W. Cohen, "Fast Effective Rule Induction," Proc. 12th Int'l Conf.Machine Learning (ICML '95), pp. 115-123, 1995.

[13] M. Dash and H. Liu, "Feature Selection for Classification,"Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.

[14] M. Dash, H. Liu, and H. Motoda, "Consistency Based FeatureSelection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery andData Mining, pp. 98-109, 2000.

[15] S. Das, "Filters, Wrappers and a Boosting-Based Hybrid forFeature Selection," Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

[16] M. Dash and H. Liu, "Consistency-Based Search in Feature Selection," Artificial Intelligence, vol. 151, nos. 1/2, pp. 155-176, 2003.

[17] J. Demsar, "Statistical Comparison of Classifiers over Multiple Data Sets," J. Machine Learning Res., vol. 7, pp. 1-30, 2006.

[18] I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification," J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.

Author's:

K.Sireesha is a student of Sree Rama Institute of Technology & Science, Kuppenakuntla,Penuballi, Khammam, TS,India.Presently she is Pursuing her M.Tech (CSE) from this college His area of interests includes Information Security, Cloud Computing, Data Communication & Networks.

S. Suresh well known author and excellent teacher. He is working as a H.O.D for the Department of M.Tech., Computer Science &Engineering, Sree Rama Institute of Technology & Science, Kuppenakuntla, Penuballi, Khammam. He has vast teaching experience in various engineering colleges. To hiscredit couple of publications both National& International conferences / journals. His area ofInterest includes Data Warehouse and Data Mining,information security, Data Communications &Networks, Software Engineering and otheradvances in Computer Applications. He has guided many projects for Engineering Students.