IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014
ISSN : 2348 - 6090
www.IJCAT.org

Feature Subset Selection Algorithm for High Dimensional Data using Fast Clustering Method

[1] A.GowriDurga, [2] A.Gowri Priya

[1] Assistant Professor(CSE), Arulmigu Meenakshi Amman College of Engineering.

[2] B.Tech[IT], Adhiparasakthi College of Engineering.

Abstract - Feature selection means finding most useful features and it will produce suitable results among entire set of features. An algorithm is used to selecting a feature and it may be evaluated from both efficiency and effectiveness point of view. Efficiency is related to the time required to find a subset of features while the effectiveness is related to quality of subset of features. Based on these, we proposed a fast clustering-based feature selection algorithm (FAST). FAST algorithm performs in two steps. First of all, features are divided into various clusters. Then the most useful feature is selected from each cluster. We adopt the minimum spanning tree (MST) to increase the efficiency of FAST. Many useful feature selection algorithms such as FCBF, Relief, CFS, Consist, FOCUS-SF are compared to FAST algorithm.

Keywords **- Feature selection, Clustering, Filter, MST.**

1. Introduction

Feature subset selection is an effective way for reducing dimensionality, eliminating irrelevant data and redundant data, increasing accuracy. There are various feature subset selection methods in machine learning applications and they are classified into four categories: Embedded, wrapper, filter and hybrid approaches.

Embedded approach is more efficient than other three approaches. Example for this approach is traditional machine learning algorithms such as decision trees and neural networks. Wrapper method gives more accuracy in learning algorithms. But here the computational complexity is large. In filter method, there is a good generality and independent of learning algorithms. But here accuracy of leaning algorithms is not guaranteed. The hybrid method is the combination of filter and wrapper method to achieve best possible performance.

We have clustered the features by graph-theoretic methods to select most representative feature related to target class. For this, we adopt MST in Fast clustering-based feature Selection algorithm (FAST). FAST algorithm performs in two steps. First of all, features are divided into various clusters. Then the most useful feature is selected from each cluster.

2. Related Works

The process of identifying and removing the irrelevant and redundant features is possible in feature is possible in feature subset selection. Due to 1) irrelevant features do not participate to the expected accuracy and 2) redundant features getting information which is already present.

Many feature subset selection algorithm can effectively removes irrelevant features but does not handle on redundant features. But our proposed FAST algorithm can remove irrelevant features by taking care of the redundant features.

In earlier days, feature subset selection has concentrate on finding for relevant features. Relief is a good example for it. But Relief is ineffective at finding redundant features. Later, Relief is extended into Relief-F to deal with noisy and incomplete data sets but it still cannot identify redundant features. CFS, FCBF, and CMIM are examples considering redundant features. FCBF is a fast filtering method that finds relevant features as well as redundancy among it. Differing from these algorithms, our proposed FAST algorithm uses the clustering-based method to choose features. It uses MST method to cluster features.

3. Feature Subset Selection Algorithm

3.1 Framework and Definitions

Feature subset selection should be able to identify and eliminate irrelevant and redundant information as possible. Because irrelevant and redundant features severely affect the accuracy of the learning machines. So we develop a

novel algorithm to deal with both irrelevant and redundant features. Finally, it will obtain a good feature subset.

Figure1 shows the feature subset selection using our proposed FAST algorithm. In this, irrelevant and redundant features are removed. To eliminate the irrelevant and redundant features, our FAST algorithm involves three steps. 1) MST is constructed 2) MST is partitioned into a forest 3) Selecting most representative features from each clusters.

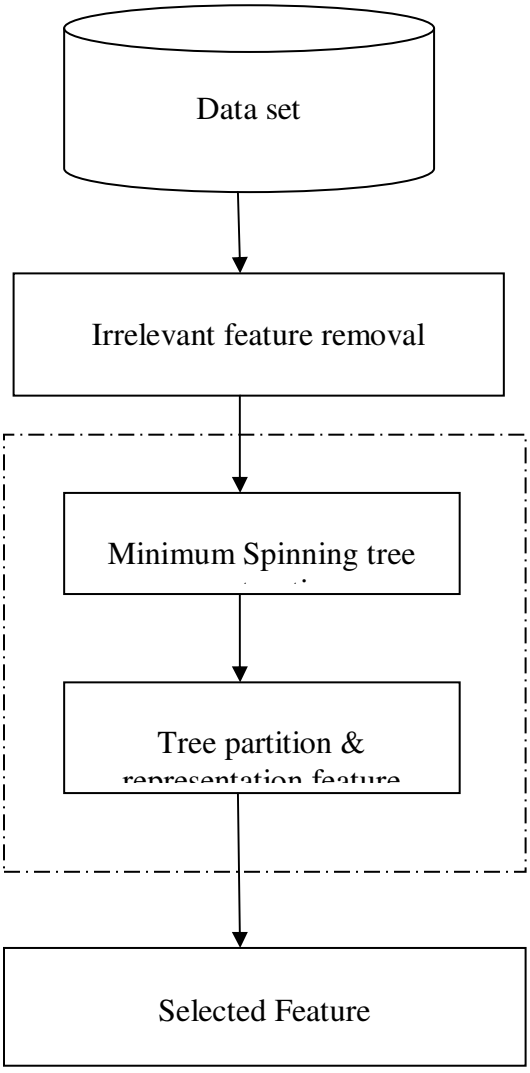

Figure1. Framework for the proposed feature subset selection algorithm

4. Modules

4.1. Distributed Clustering

In distributed clustering, words are clustered into groups using new information theoretic divisive algorithm and it is applied for text classification.

4.2. Subset Selection Algorithm

We develop a novel algorithm for feature selection to efficiently and effectively deal with both irrelevant and redundant data.

4.3. MicroArray Data

We have compared a FAST algorithm with other six algorithms. It shows that other six algorithms work well with microarray data.

4.4. Irrelevant Features

In this module we are eliminating irrelevant features by using minimum spanning tree method.

5. System Architecture

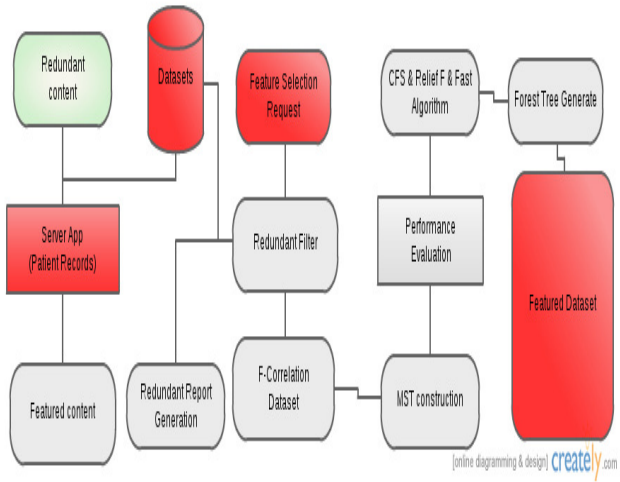

Figure2.System Architecture for patient record maintenance

Figure2 shows the system architecture diagram for patient record maintenance. In this, irrelevant and redundant features are removed and finally selected features only produced.

6. Conclusions

In this paper, we have proposed a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) eliminating irrelevant features 2) constructing MST 3) partitioning MST and selecting representative features. We have compared the performance of the proposed algorithm with five feature selection algorithm such as FCBF, Relief-F, CFS, Consist and FOCUS-SF.

References

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

[2] H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the 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 Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[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 a Feature Ranking Relieff Algorithm," Int'l J. Business Intelligence and Data Mining, vol. 4, nos. 3/4, pp. 375-390, 2009.

GowriDurga.A is an assistant professor in Department of Computer Science and Engineering at Arulmigu Meenakshi Amman College of Engineering. She received Bachelor's degree in Computer Science and Engineering from Anna University, Chennai and ME in Computer Science and Engineering from Anna University, Chennai. She has a teaching experience of one year.

Gowri Priya.A is a Final year student in Department of Information Technology at Adhiparasakthi College of Engineering.