International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 10, October 2014 ISSN 2319 - 4847

# An Improved Fast Clustering method for Feature Subset Selection on High-Dimensional Data clustering

J.K.Madhavi[1], G.Venkatesh Yadav[2]

[1] Pursuing M.Tech, Department of Computer Science, Raghu Engineering College, Dakamarri, Visakhapatnam

[2]SAsst. Professor, Department of Computer Science, Raghu Engineering College, Dakamarri, Visakhapatnam

## ABSTRACT

***In this paper, we proposed Feature extraction as the process of eliminating the irrelevant information and features during Data Mining. Feature subset selection can be analyzed as the practice of identifying and removing as lot of inappropriate and unnecessary features as achievable. This if for the reason that, irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to receiving a better analysis for that they provide typically information which is previously present in other features of all the existing feature subset selection algorithms, most of them can effectively eliminate irrelevant features but fail to handle redundant features. The improved FAST algorithm is evaluated using various types of data like text data, micro-array data and image data to represent its performance. Fast clustering algorithm work can be done in two steps. The first step is to moving out irrelevant features from the dataset, for irrelevant features are removed by the features having the value above the predefined threshold. And the second step is to eliminate the redundant features from the dataset, the redundant features is removed by constructing the Minimum Spanning Tree and separate the tree having the edge distance more than its neighbor to form the separate clusters, from the clusters features that are strongly associated with the target features are selected to form the subset of features. The Fast clustering Algorithm is more efficient than the existing feature subset selection algorithms. These can be formed in well equipped format and the time taken to retrieve the information will be short time and the Fast algorithm calculates the retrieval time of the data from the dataset. This algorithm formulates as per the data available in the dataset. By analyzing the efficiency of the proposed work and existing work, the time taken to retrieve the data will be better in the proposed by removing all the irrelevant features which gets analyzed.***

**Keywords:-** Irrelevant and redundant features, fast clustering-based feature selection algorithm, feature subset selection, Data Mining, Filter method, featured clustering, Data search, Text classification, Clustering, Rule mining

## 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. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods. Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model. In Embedded methods the process of feature selection is embedded inside the training process itself. Traditional learning algorithms like Decision trees and Artificial Neural Networks use this type of approach. The second method named Wrapper defines a possible feature subset in the target space. The results produced by mining approach are used to obtain the relevance of attributes. The intrinsic properties of training data are only taken into consideration in Filter method. A score providing the feature relevance will be calculated. The attributes with low score of relevance will be removed. The classification algorithm gets the relevant feature subset list and reduces the dimensionality of the databases by removing the redundant features. A mixture of Filter and Wrapper method forms the Hybrid method. The Hybrid method uses both the features of Filter method and the Wrapper method to select the relevant attributes using the predictive accuracy. The Filter method is considered to be the efficient one since a multidimensional dataset is reduced to a simple, fast and independent attributes list. Though Data mining process provides an immense key on useful knowledge extraction our survey focuses on the Filter method of attribute selection.

## 2. RELATED WORK

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. In the Proposed Approach, the feature subset selection algorithm the data is to be viewed as the process of identifying and eliminating as many immaterial and unneeded features as probable. This is because inappropriate features do not supply to the projecting accuracy and unneeded features do not redound to getting a better predictor for that they provide mostly information which is already present in other features. In past systems users can store the data and retrieve the data from server without any knowledge about how it is maintained and how it gets processed, but in the present system the users clearly know about the process of how to sending a request for the particular thing, and how to get a response for that request or how the system shows the results explicitly, but no one knows about the internal process of searching records from a large database. This system clearly shows how an internal process of the searching process works; this is the present advantage in our system. In text classification, the dimensionality of the feature vector is usually huge to manage. In the present system we eliminate the redundant data finding the representative data, reduce the dimensions of the feature sets, find the best set of vectors which best separate the patterns and two ways of doing feature reduction, feature selection and feature extraction. Good feature subsets contain features highly correlated with predictive of the class, yet uncorrelated with each other. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. In future the same system can be developed with the help of mobile applications to support the same procedure in mobile environments so that the users can easily manipulate their data in any place. The proposed algorithm is compared with some subset features selection algorithm. In comparison with other algorithm the fast clustering algorithm select the features are more relevant to the objective class. The feature subset algorithm removes the irrelevant features and also the redundant features. Many methods are based on searching a feature set that optimizes some evaluation function.

**Feature Selection**

Of all the existing feature subset selection algorithms, most of them can effectively eliminate irrelevant features but fail to handle redundant features. There are also algorithms that can eliminate the irrelevant features also taking care of the redundant features. Our proposed work falls into the second group. The Improved FAST algorithm is proposed to provide an efficient method to feature subset selection for different categories of data. The architecture of proposed system (Figure 2) shows the Functional diagram of our Improved FAST subset selection method. This architecture takes a search dataset as input and performs the subset selection process. The user must login with their authenticated user id and password before going in to the search process. Four types of search operations are performed here that is text search, image search, news search and sports search. Before the search operation is performed our Improved FAST performs the feature extraction process and eliminates the irrelevant and redundant information from our search space. This reduces the time complexity and increases the performance of the search. The Improved FAST eliminates irrelevant features first and from the result set it removes the redundant features. The Improved FAST method accomplishes two tasks. During the first step the features are divided into clusters using graphical cluster methods removing the irrelevant features. In the second step, the most representative feature that is closely related to the target classes is selected from each cluster to form the features subset. The efficiency of the algorithm is improved using Minimum spanning tree clustering method. The Improved FAST algorithm is evaluated using various types of data like text data, micro-array data and image data to represent its performance.

**The proposed method undergoes the Feature selection process by undergoing four phases.**

**A. Irrelevant Feature Removal**

This phase is concerned with the removal of irrelevant features that does not match with the target concept. The features are extracted as irrelevant using a match concept that reveals the relevance property between a feature and its target class. If there is no match between the values of the selected feature f and the target class c, it is said to be irrelevant and thus removed from the set of features. If the relevance measure is beyond the threshold then that feature is selected.

**B. Clustering**

The Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words or on the distribution of class labels associated with each word. As distributional clustering of words is agglomerative in nature, and result is suboptimal word clusters and acquires high computational cost. A new information-theoretic divisive algorithm for word clustering is proposed and applied it to text classification. The proposed algorithm is used to cluster features using a special metric of distance, and then makes use of the resulting cluster hierarchy to choose the most relevant attributes.

**C. Redundant Feature Removal**

The next phase in FAST method is redundant feature removal. After removing the irrelevant features, there occurs need to remove the redundant features. If a feature is embedded with redundant information, then it may not contribute to the better prediction of target classes. Redundant features completely correlate with each other. So if F is a set of features then it is said to be redundant if it has Markov Blanket within F. Assuming this as the redundant feature is removed. The major amount of work for FAST algorithm involves the computation of Symmetric Uncertainty (SU) values from which the T-Relevance and F-Correlation are calculated. This measure has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features m. The Improved FAST algorithm strives to improve the time complexity by reducing the time taken to calculate the SU values thus increasing the overall performance.

**D. Subset Selection**

Relevant features are grouped into clusters and a representative of each cluster is retrieved to get a required feature without redundancy. The 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, good feature subsets selection methods must be used to obtain features that are highly correlated with the class, yet uncorrelated with each other. A novel method is proposed which can efficiently and effectively deal with both irrelevant and redundant features, and obtains a good feature subset.

**Experimental Result**

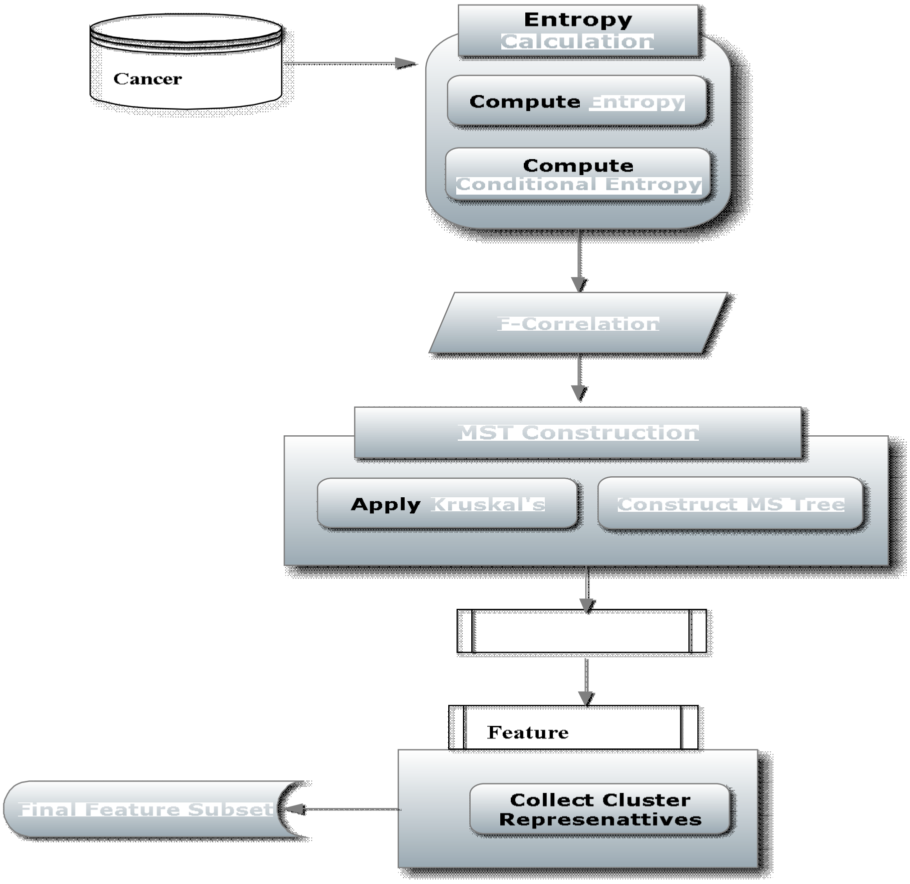
In the user module, the Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first. The Distributional clustering has been used to cluster 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, proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification, proposed to cluster features using a special metric of distance, and then makes use of the of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower. The 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, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. The major amount of work for Algorithm 1involves the computation of SU values for TR relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features m. Assuming features are selected as relevant ones in the first part, when k ¼ only one feature is selected.

**PROPOSED METHOD**

Quite different from these hierarchical clustering-based algorithms, our proposed FAST algorithm uses minimum spanning tree-based method to cluster features. Meanwhile, it does not assume that data points are grouped around centers' or separated by a regular geometric curve. Moreover, our proposed FAST does not limit to some specific types of data. 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, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other." Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. We achieve this through a new feature selection framework 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.

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters.

**System Architecture**



## Conclusion

The overall function leads to the subset selection and FAST algorithm which involves, removing irrelevant features, constructing a minimum spanning tree from relative ones (clustering) and reducing data redundancy and also it reduces time consumption during data retrieval. Thus we have presented a FAST algorithm which involves removal of relevant features and selection of datasets along with the less time to retrieve the data from the databases. The identification of relevant data's is also very easy by using subset selection algorithm. The Improved FAST algorithm is proposed to provide an efficient method for feature subset selection for different categories of data. This work is done is four phases to remove irrelevant feature, clustering similar features, removing redundant features and subset selection. These phases are applied to Data mining algorithms to reduce the number of features selected for mining. Performance based on time complexity is measured and results are shown. Thus our Improved FAST algorithm works efficiently and shows higher performance than FAST in terms of search time. In the near future this work may also be enhanced in search of video and audio files that are not performed in this work. The same set of procedures may be applied for video and audio files with large set of features. We have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

## 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.

[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 Feature Selection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.

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

## AUTHORS

**J.K.Madhavi** received her B.Tech in Computer Science Engineering from Jawaharlal Nehru Technological University, Hyderabad in 2008 and currently pursuing her M.tech in Computer Science and Engineering in Raghu Engineering College, Dakamarri, Visakhapatnam and her research interest includes Data Mining.

**G.Venkatesh Yadav** is currently working as an Assistant Professor from Department of Computer Science, Raghu Engineering College, Dakamarri, Visakhapatnam. He received his Bachelor Degree as well as Masters' Degree in Engineering and his research interest includes Data Mining, Networks.