

# Design Of Fast Clustering Based Feature Subset Selection Algorithm For High Dimensional Data

Mr. S.Sathish

Department of Computer Science, M.Phil Research Scholar, Chikkanna Govt Arts College, Tiruppur.

Email:sathishmscbdmphil@gmail.com

Mrs. B.Hemalatha

Assistant Professor& Head, PG &Research Department of Computer Science,Chikkanna Govt Arts College,

Tiruppur.Email:sabahema99@gmail.com

## ABSTRACT

A Feature selection algorithm is employed for removing irrelevant, redundant information from the data set. Amongst feature subset selection algorithm, filter methods are used because of its generality and are usually good choice when number of features are large. A Fast clustering based feature selection algorithm is based on MST method. In the FAST algorithm, features are divided into clusters by using graph-theoretic clustering method. A feature subset selection algorithm (FAST) is used to test high dimensional available image, microarray, and text data sets. Traditionally, feature subset selection research has focused on searching for relevant features. The clustering based strategy of FAST has a high probability of producing a subset of useful and independent features. In the proposed algorithm, removing of redundant data in the dataset is considered to reduce time complexity and improving learning accuracy.

**Keywords - Cluster analysis, Filter method, Graph-theoretic clustering, MST, Redundant features.**

## 1. INTRODUCTION

Data mining is a process of analyzing data and summarizes it into useful information. In order to achieve successful data mining, feature selection is an essential component. In machine learning feature selection is also known as variable selection or attributes selection. The main idea of feature selection is to choose a subset of features by eliminating irrelevant or non predictive information. It is a process of selecting a subset of original features according to specific criteria. Feature selection is an important and frequently used technique in data mining for dimension reduction. It is employed for removing irrelevant, redundant information from the data to speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility. Supervised, unsupervised and semi-supervised feature selection algorithms are developed as result of process of feature selection algorithm. A supervised feature selection algorithm determines features' relevance by evaluating their correlation with the class or their utility for achieving accurate prediction, and without labels, an unsupervised feature selection algorithm may exploit data variance or data distribution in its evaluation of features relevance and a semi-supervised feature selection algorithm uses a small amount of labeled data as additional information to improve unsupervised feature selection [2]. Feature subset selection methods can be divided into four major categories: Embedded, Wrapper, Filter, and Hybrid. The embedded method has feature selections as a part of the training process and are usually specific to given learning algorithms, and thus possibly more efficient than the other

three categories. Machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. Wrapper methods assess subsets of variables according to their relevance to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function. Filter methods are pre-processing methods. They attempt to assess the useful features from the data, ignoring the effects of the selected feature subset on the performance of the learning algorithm. Examples are methods that select variables by ranking them through compression techniques or by computing correlation with the output. 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. The important part of hybrid method is combination of filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. In the FAST algorithm, features are divided into clusters by using graph-theoretic clustering methods and then, the most representative feature that is strongly related to target classes is selected. Features in different clusters are relatively independent. A feature subset selection algorithm (FAST) is used to test high dimensional available image, microarray, and text data sets. Traditionally, feature subset selection research has focused on searching for relevant features.

## 2. Related work

Feature selection is aimed at choosing a subset of features by eliminating irrelevant or non predictive

information. It is a process of selecting a subset of original features according to specific criteria. Irrelevant features do not contribute to the accuracy and redundant features mostly provide the information which is already present in other features. There are many feature selection algorithm present, some of them are useful at removing irrelevant features but not effective to handle redundant features. Yet some of them can eliminate irrelevant features while taking care of redundant features [1]. FAST algorithm falls in to second group. One of the feature selection algorithms is Relief [3], which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is useless at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted [4]. Relief-F [5] 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. Redundant features also affect the accuracy and speed of learning algorithm; hence it is necessary to remove it. CFS [6], FCBF [7], and CMIM [9] are examples that take into consideration the redundant features. CFS [6] is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. FCBF ([7], [8]) is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. CMIM [9] iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. Different from above algorithms, FAST algorithm uses minimum spanning tree-based method to cluster features. Feature selection framework can deal with effectively and efficiently deal with irrelevant and redundant features. It is made up of 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 [1]. FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the minimum spanning tree into a forest such that each tree representing a cluster; and 3) and then the selection of representative features from the clusters. Relevant features have strong correlation with target concept hence they are always needed for a best subset, while redundant features are not needed because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally defined in terms of feature correlation and feature-target concept correlation.

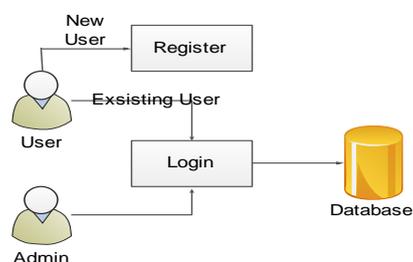
### 3. Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful

information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples. Feature subset selection is an effective way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguosness. Numerous feature subset selection methods have been planned and considered for machine learning applications. They can be separated into four major categories such as: the Wrapper, Embedded, and Filter and Hybrid methods. In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that they do not imagine that data points are clustered around centers or separated by means of a normal geometric curve and have been extensively used in tradition.

### 4. Methodology

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. The redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other features. Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some 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 has focused 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-F 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. Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other. The efficiently and effectively deal with both



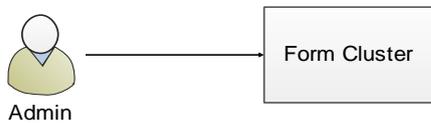
irrelevant and redundant features, and obtain a good feature subset. Fast algorithm is mainly used to obtain a good feature subset. It is a low time consuming process. Effective search is achieved based on feature search. There should be no outliers in the data. It is easy to cluster the values.

#### 4.1 User Module

In this module, Admin and 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. Authentication is providing to enhance security for data from unauthorized users. Admin maintain database by forming clustering and subset. User can access the data by searching.

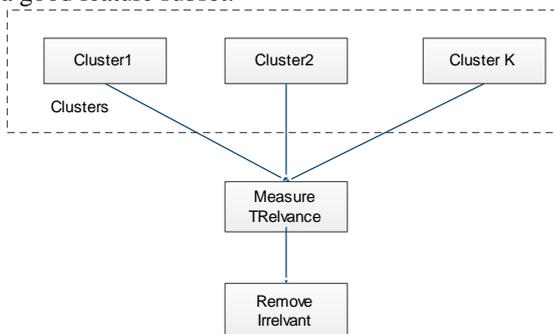
#### 4.2 Distributed Clustering

Clustering is grouping data which are having same features. In distributed clustering, words are clustered into groups using new information theoretic divisive algorithm and it is applied for text classification. The Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words.



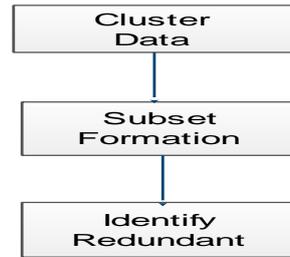
#### a) 4.3 Irrelevant feature removal

Once the right relevance measure is defined or selected, the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm T-relevance is calculated for each data. Identify the strongly related data by comparing T-relevance with the threshold value. Moreover, “good feature subsets contain features highly correlated with 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 irrelevant features, and helps to obtain a good feature subset.



#### 4.4 Subset Selection Algorithm

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. In our proposed FAST algorithm, it involves the construction of the minimum spanning tree from a weighted complete graph. Partitioning of the MST into a forest with each tree representing a cluster and the selection of representative features from the clusters. From constructed subset identify redundant data and remove it to form subset more accuracy. Redundant data are identified by measuring correlation measure between data.

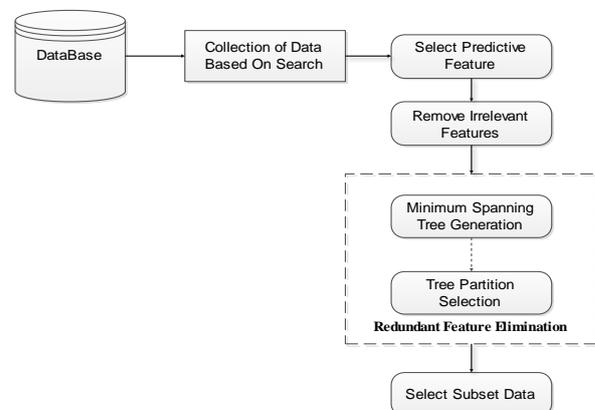


#### 4.5 Time Complexity

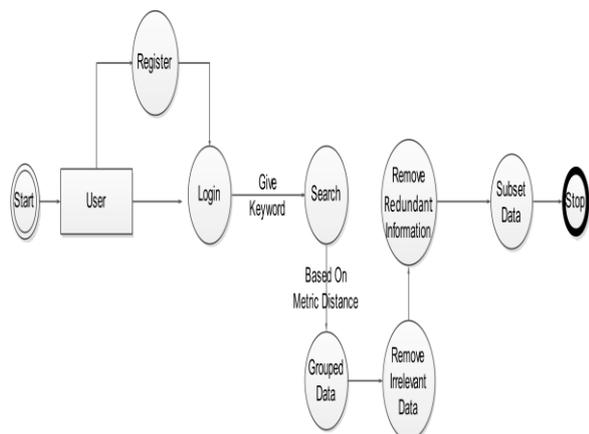
The major amount of work for FAST Algorithm involves the computation of SU (Symmetric Uncertainty) values for TR relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features  $m$ . Assuming features are selected as relevant ones in the first part, when  $k \ll m$  only one feature is selected.

#### 5. System design

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. In our proposed FAST algorithm, it involves: the construction of the minimum spanning tree from a weighted complete graph; the partitioning of the MST into a forest with each tree representing a cluster; and the selection of representative features from the clusters.



Over All DFD



## 6. Conclusion

This paper explains about the feature subset selection. It includes the modules User Module, Distributed Clustering, Subset Selection Algorithm. The FAST cluster based subset selection algorithm involves three important steps: 1. Removal of irrelevant features. 2. Elimination of Redundant features using minimum spanning tree. 3. Partitioning the MST and collect the selected features. Each cluster consists of redundant features and which is treated as single feature, so that dimensionality is reduced. A feature subset selection algorithm (FAST) is used to test high dimensional available image, microarray, and text data sets. The clustering-based strategy of FAST produces a subset of useful and independent features. The FAST algorithm can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

## References

- [1] Qinbao Song, Jingjie Ni and Guangtao Wang "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data" IEEE transactions on knowledge and data engineering vol:25 no:1 year 2013
- [2] Almuallim H. and Dietterich T.G., "Algorithms for Identifying Relevant Features", In Proceedings of the 9th Canadian Conference on AI, pp 38-45,1992.
- [3] Kira K. and Rendell L.A.," The feature selection problem: Traditional methods and a new algorithm", In Proceedings of Ninth National Conference on Artificial Intelligence, pp 129-134, 1992.
- [4] Koller D. and Sahami M., "Toward optimal feature selection", In Proceedings of International Conference on Machine Learning, pp 284-292, 1996.
- [5] Kononenko I.," Estimating Attributes: Analysis and Extensions of RELIEF", In Proceedings of the 1994 European Conference on Machine Learning, pp171-182, 1994.
- [6] Hall M.A., "Correlation-Based Feature Subset Selection for Machine Learning", Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.
- [7] Yu L. and Liu H., "Feature selection for high-dimensional data: a fast correlation-based filter solution", in

Proceedings of 20th International Conference on Machine Learning, 20(2), pp 856-863, 2003.

[8] Yu L. and Liu H., "Efficient feature selection via analysis of relevance and redundancy", Journal of Machine Learning Research, 10(5), pp 1205-1224, 2004.

[9] Fleuret F., "Fast binary feature selection with conditional mutual Information", Journal of Machine Learning Research, 5, pp 1531-1555, 2004.

[10] Hall M.A. and Smith L.A., "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper", In Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp 235-239, 1999.