An Efficient Clustering Based Irrelevant and Redundant Feature Removal for High Dimensional Data Using FAST Algorithm

Radha Saranya B.V[1], Anantha Rao Gottimukkala[2]

[1] M.Tech Student, Department of CSE, Dr Samuel George Institute of Engineering and Technology, A.P, India.

[2] Associate Professor, Department of CSE, Dr Samuel George Institute of Engineering and Technology, A.P, India.

***Abstract*—in machine learning, data mining, feature selection is the process of choosing a subset of most significant features for utilize in model construction. Using a feature selection method is that the data encloses many redundant or irrelevant features. Where redundant features are those which supply no additional information than the presently selected features, and irrelevant features offer no valuable information in any context.**

A feature selection algorithm may be expected from efficiency as well as effectiveness points of view. In the proposed work, a FAST algorithm is proposed based on these principles. FAST algorithm has various steps. In the first step, features are divided into clusters by means of graph-theoretic clustering methods. In the next step, the most representative feature that is robustly related to target classes is chosen from every cluster to make a subset of most relevant Features. Also, we use Prim's algorithm for managing large data set with effective time complexity. Our proposed algorithm also deals with the Feature interaction which is essential for effective feature selection. The majority of the existing algorithms only focus on handling irrelevant and redundant features. As a result, simply a lesser number of discriminative features are selected.

Index-Terms: **Feature Selection, Subset Selection, Feature Clustering, Graph-Based Clustering, Minimum Spanning Tree.**

I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process(KDD)),is the process of extracting samples in large data sets involving schemes at the intersection of artificial intelligence, machine learning techniques, statistics, and database systems concepts. Data mining contains six common modules of tasks those are can be known as Anomaly Recognition, Association rule mining, clustering, classification, summarization, and regression. Where as clustering is related to chore of determining groups and structures in the data that are in some way or another " related ", without using known structures in the data.

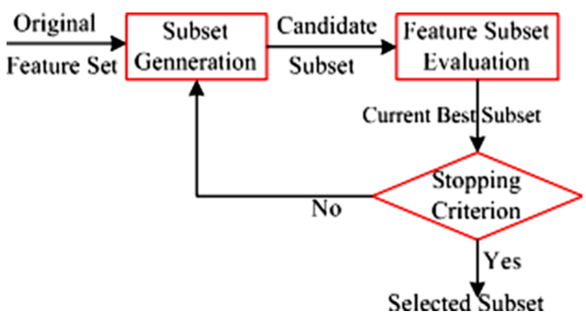
Feature selection is achieved automatically in Analysis or Examine Services, and each algorithm has a set of default techniques for wisely applying feature reduction. Feature selection is continuously performed before the model is qualified, to automatically prefer the attributes in a dataset that are most likely to be used in the model. In spite of this, you can also manually situate parameters to induce feature selection behavior. In general, feature selection works by measuring a score for each feature, and then selecting only the features that have the best scores. You can also regulate the threshold for the top scores. Analysis Services provides multiple techniques for measuring the scores, and the accurate method that is applied in any model depends on these factors:

- The algorithm used in your representation
- The data type of the feature
- Any factors that you might have set on your representation.

The choice of estimation metric heavily effect the algorithm, and it is these estimation metrics which make a distinction between the three main categories of feature selection algorithms: wrappers, filters and embedded methods. Wrapper techniques use a predictive model to score feature subsets. Each recent subset is used to train a model, which is experienced on a hold-out set. Counting the amount of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. Filters are mostly fewer computationally exhaustive than wrappers, but

ISSN: 2278 – 7798

they generate a feature set which is not regulated to a specific type of predictive model. Many filters supply a feature ranking relatively than an explicit best feature subset, and the interrupt point in the ranking is chosen via cross-validation. The wrapper methods use the predictive exactness of a determined learning algorithm to determine the integrity of the selected subsets, the accuracy of the machine learning algorithms are generally high. However, the majority of the selected features are partial and the computational complexity is huge. Through the filter feature selection methods, the function of cluster analysis has been demonstrated to be extra valuable than traditional feature selection algorithms.



In this proposed work. We have clustered the features by using the graph-theoretic approach to select most representative feature related to target class. To do that, we take on Minimum-Spanning-Tree (MST) in Fast clustering-bAsed feature Selection algoriThm (FAST). FAST algorithm completes in two steps. First of all, features are separated into various clusters. After that the most effective feature is selected from each cluster.

II PROBLEM STATEMENT

PREVIOUS SYSTEM:

The process of detecting and eliminating the irrelevant and redundant features is possible in feature is possible in feature subset selection. Because of 1) irrelevant features do not involve to the expected accuracy and 2) redundant features getting information which is previously exists.

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

In former days, feature subset selection has focused on discovering for relevant features. Relief is an excellent example for it. But Relief is unsuccessful at discovering redundant features. Later on, Relief is extended to Relief-F to contract with noisy and partial data sets but even now it cannot discover redundant features. CFS, FCBF, and CMIM are the examples of considering redundant features. FCBF is a fast filtering technique that discovers relevant features plus redundancy among it. Conflicting from these algorithms, proposed FAST algorithm uses the clustering-based scheme to select features. It uses MST (Minimum Spanning Tree) technique to cluster the features.

Disadvantages

1. the most part of the selected features are lesser and the computational complexity is huge.
2. Their computational complexity is low, although the correctness of the learning algorithms is not sure.

PROPOSED SYSTEM:

According to the Previous System, Inappropriate features, along with unnecessary features, severely influence the accuracy of the learning machines. Hence, feature subset selection algorithm should be capable to discover and eliminate as much of the inappropriate and unnecessary information as possible. Additionally, "superior feature subsets include features highly interrelated with (predictive of) the class, still uncorrelated with (not predictive of) each other."

For above mentioned problem, we develop a unique algorithm which can proficiently and successfully deal with both inappropriate and unnecessary features, and acquire a good feature subset. We achieve this via a recent feature selection framework (presented in Fig.1) which composed of the two associated mechanism of inappropriate feature removal and unnecessary feature removal. The previous gained features related to the intention concept by removing inappropriate ones, and the later removes unnecessary features from related ones through selecting representatives from unlike feature

ISSN: 2278 – 7798

clusters, and therefore constructs the final relevant subset. Our Proposed FAST algorithm, has some different steps (i) the Production of the minimum spanning tree (MST) from a weighted complete graph; (ii) the division of the MST into a forest with each tree will represents a cluster; and (iii) the final selection of strongly related features from the clusters

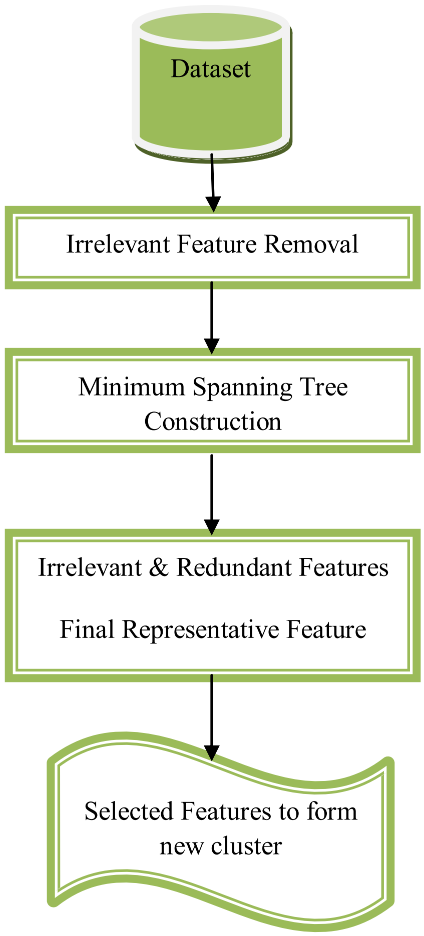


Fig1. Structure for feature subset selection algorithm

III SYSTEM DEVELOPMENT

- Load the Dataset
- Generate Subset Partition
- Irrelevant Feature Removal
- MST Construction
- Selected Features List & Centroid Based Clustering

Load the Dataset:

According to the system architecture, First Load the dataset into the process. The dataset has to be preprocessed for eliminating absent values, noise and outliers. Then the given dataset should be transformed into the arff format which is the standard format for WEKA toolkit.

Generate Subset Partition:

Generate Partition is the step to divide the whole dataset into partitions, will be able to classify and identify for irrelevant & redundant features. A strategy for reviewing the quality of model simplification is to division the data source.

Irrelevant Feature Removal:

Useful way for sinking dimensionality, eliminating inappropriate data, rising learning accuracy, and civilizing result comprehensibility. The inappropriate feature removal is directly once the right relevance quantify is defined or chosen, while the unnecessary feature elimination is a bit of complicated.

We firstly present the symmetric uncertainty (SU), Symmetric uncertainty pleasures a couple of variables symmetrically, it give back for information gain's bias in the direction of variables with greater values and regulates its value to the range [0, 1]

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)}.$$

MST Construction:

This can be shown by an example, suppose the Minimum Spanning Tree shown in Fig.2 is produced from a complete graph G. In organize to cluster the features, we first go across all the six edges, and then decided to remove the edge ($F0$, $F4$) because its weight ($F0$,4)=0.3 is lesser than both SU($F0$,C)=0.5andSU($F4$,C)=0.7. This constructs the MST is grouped into two clusters denoted as ($T1$) and ($T2$).To generate MST we used Prim's Algorithm.

ISSN: 2278 – 7798

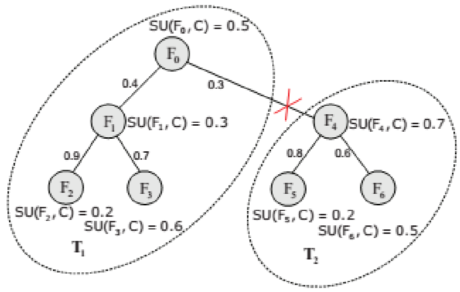


Fig 2. Clustering Step

Selected Features List & Centroid Clustering:

Ultimately it includes for final feature subset. Then calculate the accurate/relevant feature. These Features are relevant and most useful from the entire set of dataset. In centroid-based clustering method, clusters are denoted by a central vector, which might not essentially be a member of the data set. While the number of clusters is fixed to k, *k*-means clustering gives a proper definition as an optimization trouble: find the k cluster centers and allocate the objects to the nearest cluster center, such that the four-sided figure distances from the cluster are minimized.

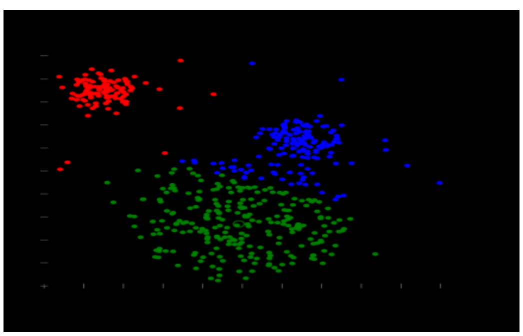


Fig 3. Centroid Clustering using K-Means

FAST Algorithm:

Algorithm 1: FAST

inputs: $D(F_1, F_2, ..., F_m, C)$ - the given data set
θ - the T-Relevance threshold.

output: S - selected feature subset .

//==== Part 1 : Irrelevant Feature Removal ====

for $i = 1$ *to* m **do**
 T-Relevance = SU (F_i, C)
 if T-Relevance $> \theta$ **then**
  $S = S \cup \{F_i\}$;

//==== Part 2 : Minimum Spanning Tree Construction ====

G = NULL; *//G is a complete graph*
for *each pair of features* $\{F'_i, F'_j\} \subset S$ **do**
 F-Correlation = SU (F'_i, F'_j)
 Add F'_i *and/or* F'_j *to* G *with* F-Correlation *as the weight of the corresponding edge;*
minSpanTree = Prim (G); *//Using Prim Algorithm to generate the minimum spanning tree*

//==== Part 3 : Tree Partition and Representative Feature Selection ====

Forest = minSpanTree
for *each edge* $E_{ij} \in$ Forest **do**
 if $\mathrm{SU}(F'_i, F'_j) < \mathrm{SU}(F'_i, C) \wedge \mathrm{SU}(F'_i, F'_j) < \mathrm{SU}(F'_j, C)$ **then**
  Forest = Forest $- E_{ij}$

$S = \phi$
for *each tree* $T_i \in$ Forest **do**
 $F^j_R = \mathrm{argmax}_{F'_k \in T_i} \mathrm{SU}(F'_k, C)$
 $S = S \cup \{F^j_R\}$;
return S

IV SCREEN SHOTS

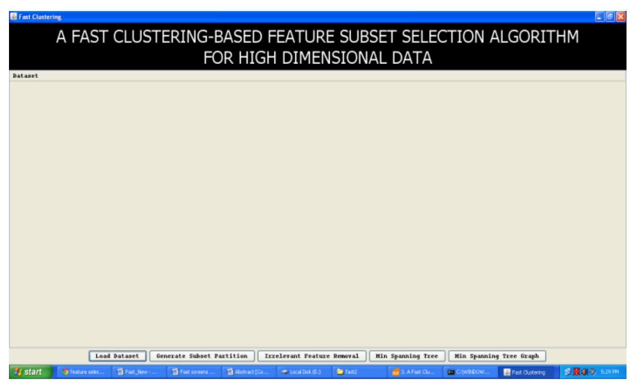


Fig 4. Main Application used to load the dataset

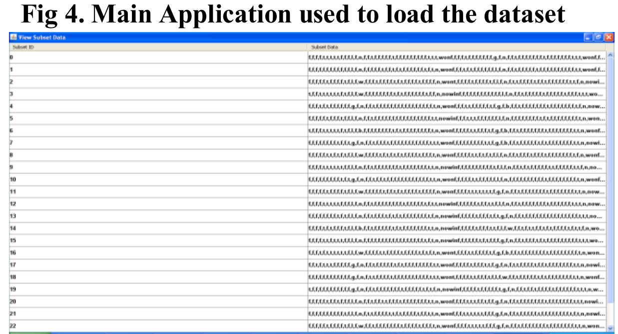


Fig 5. Generated partitioned subset data

ISSN: 2278 – 7798

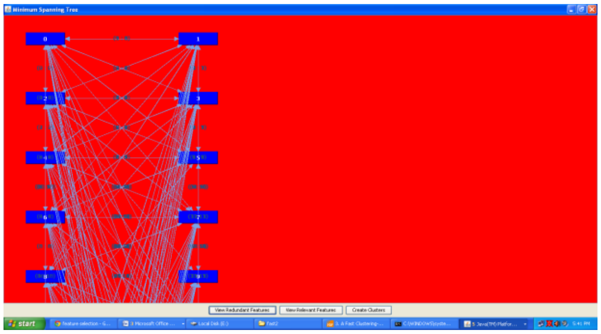


Fig 6. Constructed MST

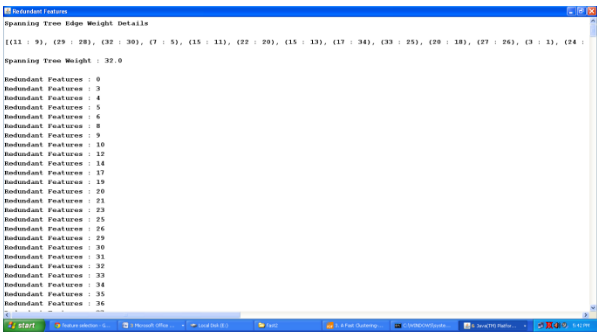


Fig 7. After removal of Redundant Features

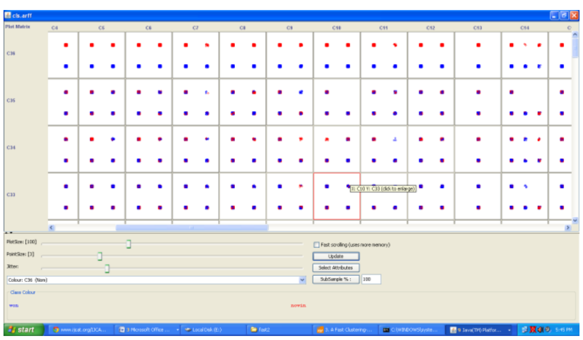


Fig 8. Selected Features formed as new cluster

V CONCLUSION

In this paper, we presented a clustering-based feature subset selection algorithm for high dimensional data. The algorithm includes (1) eliminating irrelevant features, (2) producing a minimum spanning tree from relative ones, and (3) partitioning the MST and selecting most useful features. (4) Formed as new clusters from the selected most useful features. In the proposed algorithm, a cluster consists of features. Every cluster is treated as a distinct feature and thus dimensionality is radically reduced.To extend this work, we are planning to explore various categories of correlation measures, and study some proper properties of feature space.

VI REFERENCES

[1] Dash M. and Liu H., Consistency-based search in feature selection. Artificial Intelligence, 151(1-2), pp 155-176, 2003.

[2] Hall M.A., Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000.

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

[4] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," July1994.

[5] M. Last, A. Kandel, and O. Maimon, "Information-Theoretic Algorithm for Feature Selection," Pattern Recognition Letters, vol. 22, nos. 6/7, pp. 799-811, 2001.

[6] 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.

[7] Raman B. and Ioerger T.R., Instance-Based Filter for Feature Selection, Journal of Machine Learning Research, 1, pp 1-23, 2002.

Anantha Rao Gottimukkala received B.Tech (CSE) Degree from JNT University in 2007 and M.Tech (SE) Degree from JNTUK Kakinada in 2009. He has 06+ years of teaching experience. I Worked as Asst.Prof. in 2008-2011 at Kakinada institute of Engineering & Technology[KIET],Korangi, Kakinada. He joined as Assistant Professor in Dr.Samuel George Institute of Engineering & Technology, Markapur, India in 2011 till to. Presently he is working as Associate Professor in CSE Dept. His Interested research areas are Image Processing, Algorithms', Computer Networks. He attended Various National and International Workshops and Conferences.

Radha Saranya B.V, pursuing her M.tech in computer science from Dr.Samuel George Institute of Engineering and Technology, Markapur, Prakasam District, A.P, India. Affiliated to Jawaharlal Nehru Technological University, Kakinada. Approved by AICTE, NEW DELHI.

ISSN: 2278 – 7798