Enhanced Mining of High Dimensional Data Using Efficient Fast Clustering Algorithm

P . Lakshmi Reddy
Dept. of CSE,
Sree Vidyanikethan Engg. College,
Tirupati, A.P.

Mr . Shaik Salam
Associate Professor, Dept. of CSE,
Sree Vidyanikethan Engg. College,
Tirupati, A.P.

Dr . T . V . Rao
Head, Dept. of CSE,
PVP Siddhartha Inst.of Tech.,
Vijayawada, A.P.

***Abstract*—The concept of feature selection involves identification of a subset of the most useful features that produce compatible results similar to the original entire set of features. A feature selection algorithm may be evaluated from both efficiency and effectiveness points of view. Efficiency involves the time required to find a subset of features, whereas effectiveness involves the quality of the subset of features. The FAST algorithm works in two steps. Features are divided into clusters through graph-theoretic clustering methods in the first step and the most representative feature that is strongly related to the target class is selected from each cluster to form a subset of features in the second step. The main contribution of the Relief is considered to be one of the most successful algorithms for evaluating the quality of features.It is used for efficiently estimating feature quality. The algorithm holds a weight vector over all features and updates the vector according to the sample points that are presented earlier. The Relief-F algorithm extends the Relief algorithm, by finding the near hit and near miss using the Manhattan(L1)norm rather than the Euclidian(L2) norm.**

Index Terms*—Feature subset selection, filter method, feature clustering, graph-based clustering, Relief and Relief-F algorithms

I. INRTODUCTION

Data mining can be referred to a process of discovering knowledge from data with the main emphasis on uncovering interesting data patterns that are being hidden in large data sets. Cluster analysis can be referred to a process of grouping a set of objects into various classes of similar objects. Data objects are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Feature subset selection is the process of identifying and removing irrelevant and redundant features as many as possible. This is because 1) irrelevant features do not contribute to predictive accuracy, and 2) redundant features do not rebound in getting a better predictor and they provide mostly the informationwhich is already present in other feature(s). Among the many feature subset selection algorithms, some of them effectively eliminate irrelevant features but fail to handle redundant features and some others eliminate the irrelevant while taking care of the redundant features. The proposed FAST algorithm falls into the second group. Among the feature selection algorithms, the Relief algorithm is considered to be one of the most successful ones owing to its simplicity and effectiveness. The key idea of Relief is to iteratively estimate feature weights based on its ability todiscriminate the neighbouring models. Relief-F was the extended Relief algorithm to handle noisy and missing data which solves multiclassification issues, where the original Relief algorithm cannot deal with them. Feature selection is a process to select a subset of original features. The optimality of a feature subset is measured by an evaluation criterion. As the dimensionality of a domain expands, the number of features also increases. Finding an optimal feature subset among feature subsets is usually intractable and many of such problems have been shown to be of NP-hard type. A typical feature selection process consists of four basic steps as shown in Fig. 1, namely, subset generation, subset evaluation, stopping criterion, result validation. Subset generation is a procedure to produce candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one based on certain evaluation criteria. If the new subset turns out to be better, the previous best subset is replaced by it. The process of subset generation and evaluation is performed until a given stopping criterion is satisfied iteratively. The selected best subset usually needs to be validated by prior knowledge or different tests through synthetic and/or realworld data sets. Feature selection can be found in many areas of data mining such as classification, clustering, association rules, and regression. Feature selection is also referred to as subset or variable selection in Statistics.

II. FEATURE SUBSET SELECTION ALGORITHM

The FAST algorithm deals with the irrelevant features, along with redundant features, since it severely affects the accuracy of the learning machines. Thus, feature subset selection should identify and remove as much of the irrelevant and redundant information as possible. Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other within the class.

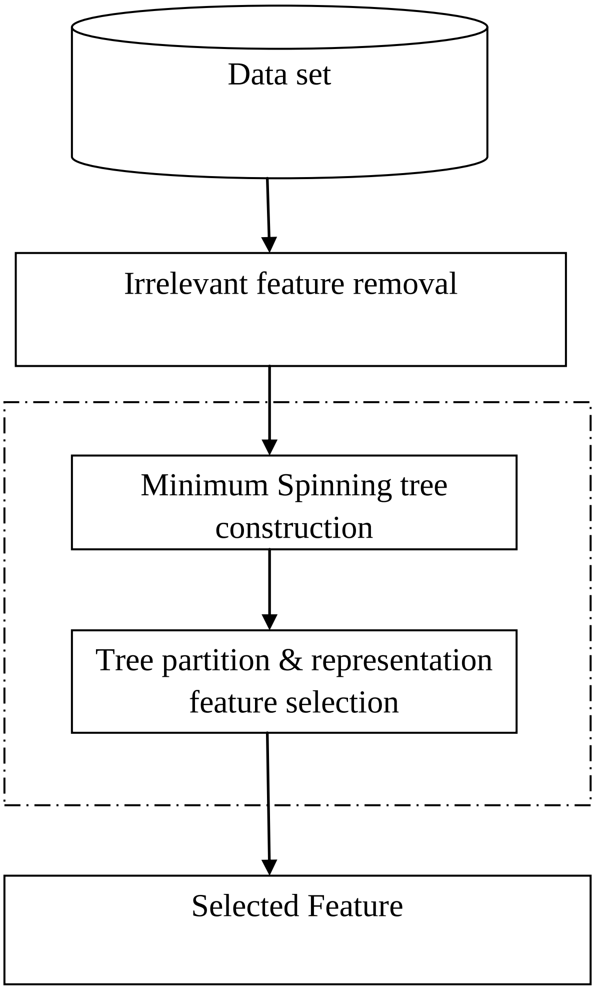
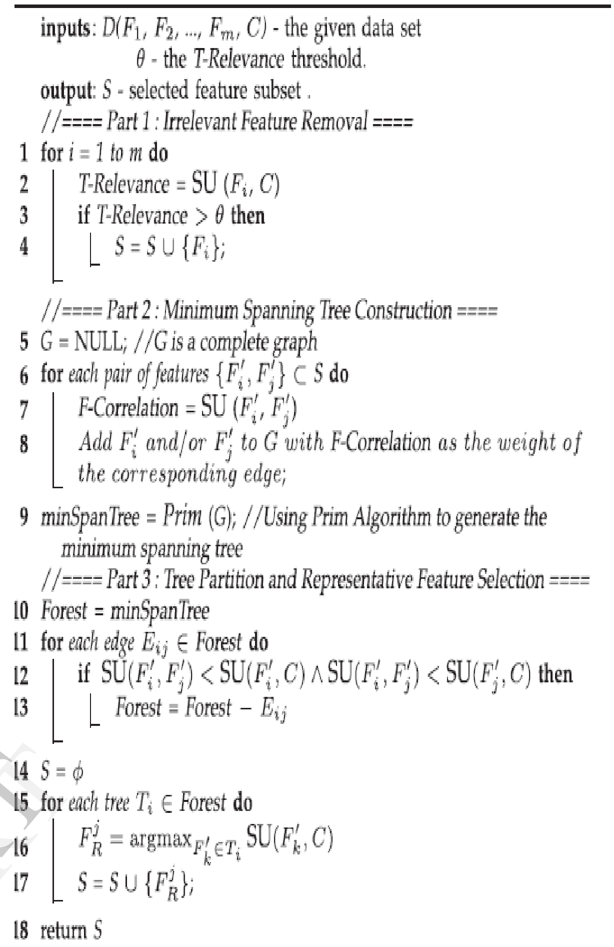

Data set
Irrelevant feature removal
Minimum Spinning tree construction
Tree partition & representation feature selection
Selected Feature

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

A new feature selection framework is shown in Fig. 1 which is composed of the two connectedcomponents i.e., irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target by eliminating the irrelevant ones, and the latter removes redundant features from relevant ones by choosing representatives from different feature clusters, and thus gives the final subset. The irrelevant feature removal is evident once the right relevance measure is defined or selected, while the redundant feature elimination is a bit complicated. The proposed FAST algorithm involves the 1) construction of minimum spanning tree from a weighted complete graph, 2) partitioning of Minimum Spanning Tree into a forest with each tree representing a cluster and 3) selection of representative features from the above represented clusters.

Algorithm 1. FAST



```
inputs: $D(F_1, F_2, ..., F_m, C)$ - the given data set
            $\theta$ - 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 $S$
```

III. RELIEF ALGORITHM

Features and models are represented by means of feature weight value, and there are still shortcomings in Relief algorithm, e.g., when all model types involved in the present problem are already definite, then there are certain features that still include certain model types which are not referred in the current issue. In this case, these features, which are straightaway substituted in Relief, are considered to have an intimate relationship with the model types, regardless of whether they are related to model types.Therefore, Relief sometimes performs blind selection strategy, which is not expected to occur. In the following, a thorough interpretation of blind selection is provided.

The procedure of Relief algorithm is given here. In each iteration, an instance x is randomlyselected and then the two nearest neighbors of x are found, one from the same classification which is termed as the *nearest hit* or NH and the other from a dissimilar classification which is termed as the *nearest miss* or NM. The weight of thei^{th} feature is then updated:

The Relief algorithm was actually designed to deal with binary problems whereas Relief-F was proposed to dispose multiclass problems by the weight update rule.

The major drawbacks of Relief are defined as follows:

Definition 1: If feature η contains one or more model types, and if the model type space does not include the problem to be resolved, we designate η as a bogus feature.

Definition 2: A model type, which does not exist in model type space in real applications and isrepresented by bogus features, is defined as connotative classifications (CC). When it is compared with other features, bogus features usually perform some special functionality which can besummarized as below:

- Regardless of whether there is strong correlation between bogus features and model type, the weights of bogus features achieve a larger value. Accordingly, the bogus feature is regarded as an informative feature which has a remarkable correlation with model type.
- Adopted a feature subset comprising ofa bogus feature, pattern recognition will deterioratethe classification performance.

Bogus feature reveals the distribution of instance set offeature η. The instance x can be accurately distinguished between model *A class* and model *B class*. Model *C class* and model *D class* are the unexpected model types which need notbe transacted in this case, and then feature η possesses the habit of abogus feature.

IV. CONCLUSIONS

This paper presents a FAST clustering algorithm. By using this algorithm we can search the feature subsets in parallel, by using Relief algorithm which in turn is used for reducing irrelevant features and redundancies and also provides the efficiency and effectiveness of the feature subsets.

REFERENCES

[1] Qinbao Song, Jingjie Ni, and Guangtao Wang, “A Fast Clustering
Based-Feature Selection Algorithm for High-Dimensional Data”,
IEEE transactions on knowledge and data engineering, vol. 25, no.
1, January 2013.

[2] K. Kira and L.A. Rendell, “The Feature Selection Problem
Traditional Methods and a New Algorithm,” Proc. 10th Nat’l Conf.
Artificial Intelligence, pp. 129-134, 1992.

[3] D. Koller and M. Sahami, “Toward Optimal Feature
Selection,” Proc. Int’l Conf. Machine Learning, pp. 284-292, 1996.

[4] I. Kononenko, “Estimating Attributes: Analysis and
Extensions of RELIEF,” Proc. European Conf. Machine
Learning, pp. 171-182, 1994.