ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

A Relevant Clustering Algorithm for High-Dimensional Data

Bini Tofflin.R[1], Kamala Malar.M[2], Sivasakthi.S[3]

M.Tech/Information Technology, Dr.Sivanthi Aditanar College of Engineering, Tiruchendur, Tamilnadu, India[1, 3]

Asst. prof/Information Technology, Dr.Sivanthi Aditanar College of Engineering, Tiruchendur, Tamilnadu, India[2]

ABSTRACT: Clustering is widely used data mining model that partitions data points into a set of groups, each of which is called a cluster. With the emerging growth of computational biology and e-commerce applications, high-dimensional data becomes very common. Thus, mining high-dimensional data is more needed. There are some main challenges for mining data of high dimensions, such as the curse of dimensionality and more crucial, the meaningfulness of the similarity measure in the high dimension space. The main goal of feature selection is to select a subset of useful features. The relevant features of subset are selected correctly and then the entire set gives accurate results. For this reason, feature subset selection is used in the high-dimensional data. The good subsets of features are selected by using feature selection method. The feature selection methods are mainly used in application of learning algorithms. Feature selection methods decrease the dimensionality of the data and allow learning algorithms to operate efficiently and more effectively. The proposed Relevant Clustering Algorithm is used for finding the subset of features. A Relevant clustering algorithm renders efficiency and effectiveness to find the subset of features. Relevant clustering algorithm work can be done in three steps. First step elimination of irrelevant features from the dataset; the relevant features are selected by the features having the value greater than the predefined threshold. In the second step selected relevant features are used to generate the graph, divide the features using graph theoretic method, and then clusters are formed by using Minimum Spanning Tree. In the third step find the subsets features that are more related to the target class is selected. The Relevant Clustering Algorithm is more efficient than the existing features subset selection algorithms RELIEF, FCBF, CFS, FOCUS-2 and INTERACT.

KEYWORDS: Feature selection, feature relevance, subset feature, minimum spanning tree, feature redundancy.

I. INTRODUCTION

Clustering method is used for grouping together data's that are similar to each other and dissimilar to the data's belonging to other clusters. Feature selection, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity [1]. A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only two to three dimensions. Human eyes are good at judging the quality of clustering for up to three dimensions. Finding clusters of data objects in high-dimensional space is challenging, especially considering that such data can be sparse and highly skewed. Irrelevant features provide no useful data's related to the target classes. The irrelevant data removal increase learning accuracy of a machine.

The feature selection is a process of finding a subset of original features so that the feature space is reduced. The objectives of feature selection include: building simpler and more comprehensible models, improving data mining performance, and helping prepare, clean, and understand data. The Feature selection technique is used in the data contains many replicated and irrelevant features. Feature selection is an essential step in successful data mining applications, which can effectively reduce data dimensionality by removing the irrelevant (and the redundant) features.

Feature subset selection is the process used for moving out the irrelevant and redundant information. 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 four main categories of methods used in feature selection algorithms are wrappers, filters, embedded and hybrid methods. The wrapper method [6] requires predefined learning algorithm to select the features. Wrapper methods use a predictive model to score feature subsets. As wrapper methods train a new model for each subset, they are very computationally intensive. The wrapper methods are beneficial but require more time to select the features. The selected features are limited and computational complexity is large in wrapper methods.

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, whilst still capturing the usefulness of the feature set. Filters methods have proven to be much efficient than wrappers in selecting the subset of features. The filter methods are used when the numbers of features are large and it is computationally effective. The embedded methods[2][7] is more efficient than the other methods In the embedded methods use feature selection as a function of the training process and are usually specific to given learning algorithms. The hybrid methods[8] are combinations both feature methods they are filter and wrapper methods to achieve best possible performance on high dimensional data with similar time complexity of filter algorithms. The filter methods are usually beneficial when the number of features is very large. Thus, the filter method is used in this paper.

In this paper clustering of data to form subset of features. The proposed algorithm is Relevant clustering algorithm consist of three steps for selecting the subset of features from the dataset. First, step elimination of irrelevant features from the dataset; the irrelevant features are removed by the features having the value less than the predefined threshold. In the second step relevant features are used to compute the graph, separate the features using graph theoretic method, and then clusters are formed by using Minimum Spanning Tree. In the third step features that are more related to the target class are selected from the each cluster to form the subsets features.

II. RELATED WORKS

The proposed algorithm is compared with some subset features selection algorithm. In comparison with other algorithm the Relevant clustering algorithm select the features are more relevant to the target class. In Relief feature selection algorithm is a well known and good feature set estimator. RELIEF algorithm [3] that uses instance based learning to assign a relevance weight to each feature. Each feature's weight based on its ability and then finds the relevant features. Features are ranked by weight and those that exceed a user-specified threshold are selected to form the final subset. A feature is relevant if its weight of relevance is greater than a threshold value. The feature set measure proposed by RELIEF algorithm is efficient than the wrapper methods. The algorithm is simple, easy to implement and computationally efficient. The computational cost is low in this algorithm. The RELIEF algorithm removes irrelevant features but does not able to remove the redundant features.

The INTERACT feature subset algorithm [16] used for finding the interact features then remove the irrelevant features, the redundant features. This algorithm selected feature is relevant if it is strongly correlated with the target concept. INTERACT Algorithm avoids the repeatable scanning of datasets. This algorithm achieves good time performance. A Fast Binary feature selection technique based on Conditional Mutual Information Maximization [5]. The Conditional Mutual Information Maximization criterion (CMIM) does not select a feature similar to already picked ones. The CMIM select features that is individually powerful and features provide information about the class to predict. The CMIM algorithm does not require the tuning of any parameter.

The FOCUS-2 algorithm [11]that exactly implements the MIN FEATURES bias. The FOCUS-2 algorithm keep features in the queue those that may contain a solution. Initially, the queue contains only the element which represents the whole power set. In each iteration the queue is partitioned into disjoint subspaces, and those subspaces that cannot contain solutions are pruned from the search. The FOCUS-2 only prunes subspaces that cannot be complete, and it will not miss any sufficient feature subsets. The number of sufficiency tests performed by FOCUS-2 will typically be much less.

Fast Correlation Based Filter (FCBF) Algorithm [15] is a fast filter method .FCBF algorithm finding a set of predominant feature FCBF algorithm can identify relevant features as well as replicated features. FCBF algorithm calculates the SU value for each feature selects relevant features list based on the predefined threshold and then processes the ordered list to remove redundant features. The FCBF algorithm only keeps predominant ones among all the selected relevant features. All features contain numerical values calculations are used in the FCBF algorithm.

FCBF algorithm does not provide the correlation between relevant features. Correlation based Feature Subset Selection (CFS) Algorithm[13] selects the features highly correlated with the target. CFS explores the search space using the Best First search. Merits of CFS are it does not need to reserve any part of the training data for evaluation purpose and works well on smaller data sets. It selects the maximum relevant feature and avoids the re-introduction of redundancy. But the drawback is that CFS cannot handle problems where the class is numeric.

III. PROPOSED WORK

The main goal of proposed Relevant Clustering Algorithm is to remove the irrelevant features and redundant features to form the subsets of features. The feature selection algorithms are to remove the irrelevant features and the redundant features. Feature selection is used mainly for the data analysis process. Subset selection evaluates a subset of features as a group for suitability. Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. 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. The Relevant clustering algorithm comprised of three steps. First step find the relevant features, second step is to eliminate redundant features and generate the clusters using graph theoretic clustering methods, and third step the feature that strongly related to the target class is selected from each cluster to form subset of features.

3.1 RELEVANT Clustering Algorithm Framework

The figure.1 shows the proposed RELEVANT clustering algorithm framework which is used to select the subset features. From the given data set the features are selected and the features are compared with the predefined threshold if the feature having the value below the threshold is removed as the irrelevant features. Relevant features have strong correlation with target concept. The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features. Therefore, symmetric uncertainty as the measure of correlation between either two features or a feature and the target concept. The symmetric uncertainty values are calculated for each feature in the data set. The features whose SU values are greater than a predefined threshold comprise the target relevant feature subset. Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines.

The relevant Features are divided into clusters by using graph-theoretic clustering methods along with Minimum Spanning Tree (MST). The general graph-theoretic clustering is simple: compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter than its neighbors. Thus for graph G build an MST, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well-known Prim algorithm. A Minimum Spanning Tree (MST) in an undirected connected weighted graph is a spanning tree of minimum weight (among all spanning trees). A Minimum Spanning Tree use Prim's algorithm. The Prim's algorithm makes a nature choice of the cut in each iteration and it grows a single tree and adds a light edge in each iteration.

For the relevant features form the graph G and set edge weight value then delete any edge greater than its neighbors and the outcome form the forest then the Minimum spanning tree clustering algorithm used to form the clusters. Each cluster is independent of each other. From the clusters generate the subset of features that are more related to the objective class. Each feature in the generated clusters has different features. The features that are relevant to the target classes are selected from each cluster. The selected features from the clusters form the Feature Subset.

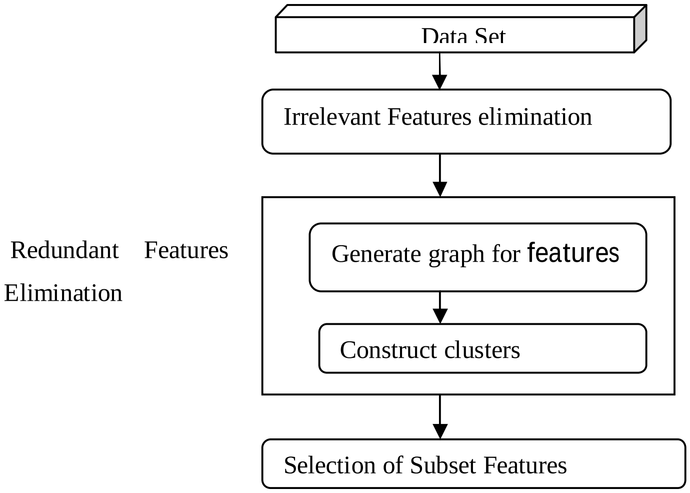
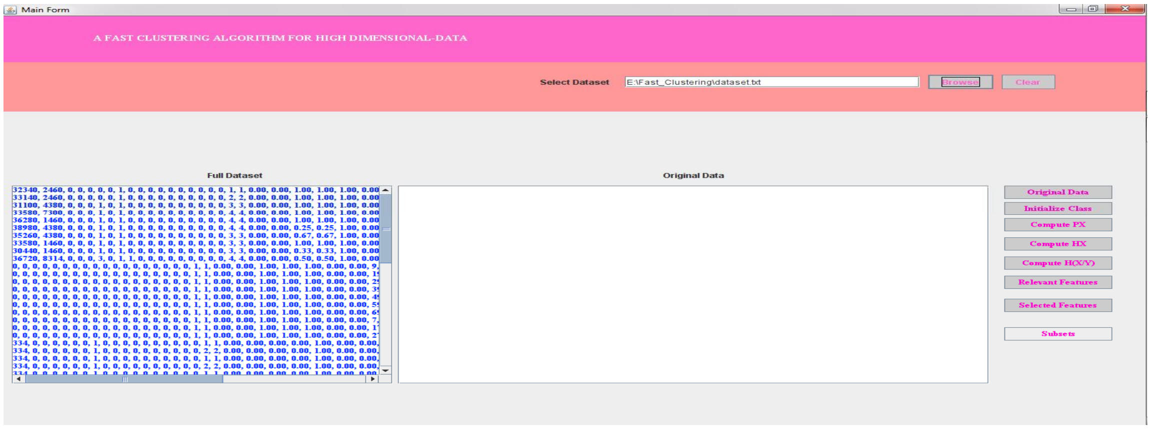

Data Set
Irrelevant Features elimination
Redundant Features Elimination
Generate graph for features
Construct clusters
Selection of Subset Features

Figure 1.Proposed Algorithm Framework

IV. RESULTS AND DISCUSSIONS

The results of paper will be briefly discussed with its simulated output. The proposed RELEVANT algorithm, it involves removal of irrelevant features from the attacker dataset, 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 to form five subset of features

4.1. Attacker Dataset



A FAST CLUSTERING ALGORITHM FOR HIGH DIMENSIONAL-DATA

Select Dataset E:\Fast_Clustering\dataset.txt

Full Dataset

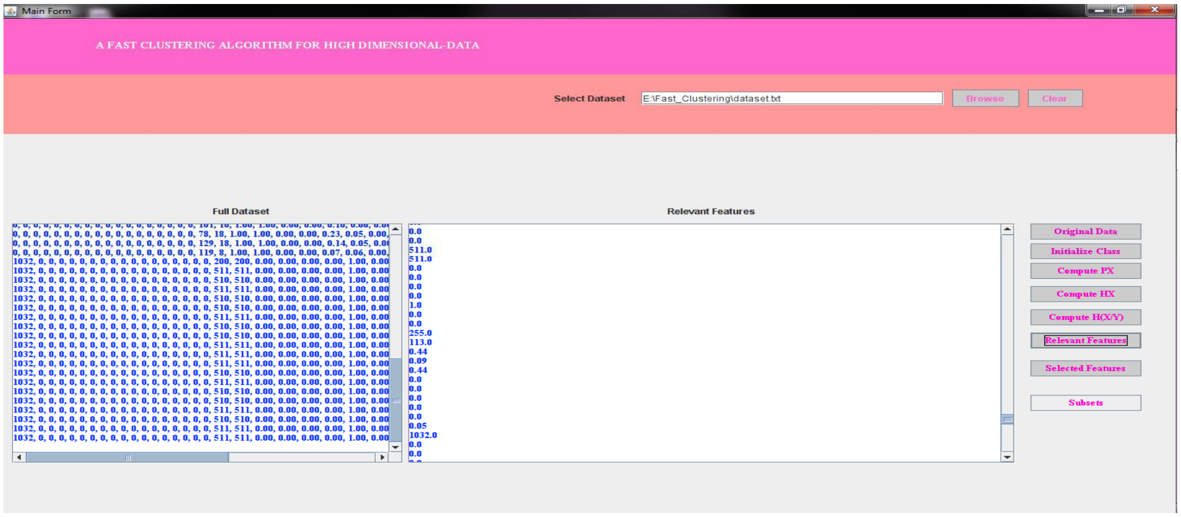
```
32340, 2460, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00
33140, 2460, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00
31100, 4380, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00
33580, 7300, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 4, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00
36280, 1460, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 4, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00
38980, 4380, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 4, 0.00, 0.00, 0.25, 0.25, 1.00, 0.00
35260, 4380, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 0.00, 0.00, 0.67, 0.67, 1.00, 0.00
33580, 1460, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00
30440, 1460, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 0.00, 0.00, 0.33, 0.33, 1.00, 0.00
36720, 8314, 0, 0, 0, 3, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 4, 0.00, 0.00, 0.50, 0.50, 1.00, 0.00
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 9,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 1
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 2
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 3
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 4
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 5
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 6
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 7,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 1
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 1.00, 1.00, 1.00, 0.00, 0.00, 2
334, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00,
334, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00,
334, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00,
334, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00,
334, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00,
```

Original Data

Figure 4.1: Attacker Dataset

In this Figure 4.1 shows the full dataset consist of attacker features. In the full dataset five groups of attacker features are available; the RELEVANT clustering algorithm is used to find the subsets of attacker features.

4.2 Selection of Relevant Features



Select Dataset E:\Fast_Clustering\dataset.txt

Full Dataset

```
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 101, 10, 1.00, 1.00, 0.00, 0.00, 0.10, 0.00, 0.0
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 78, 18, 1.00, 1.00, 0.00, 0.00, 0.23, 0.05, 0.00,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 129, 18, 1.00, 1.00, 0.00, 0.00, 0.14, 0.05, 0.0
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 119, 8, 1.00, 1.00, 0.00, 0.00, 0.07, 0.06, 0.00,
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 200, 200, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
```

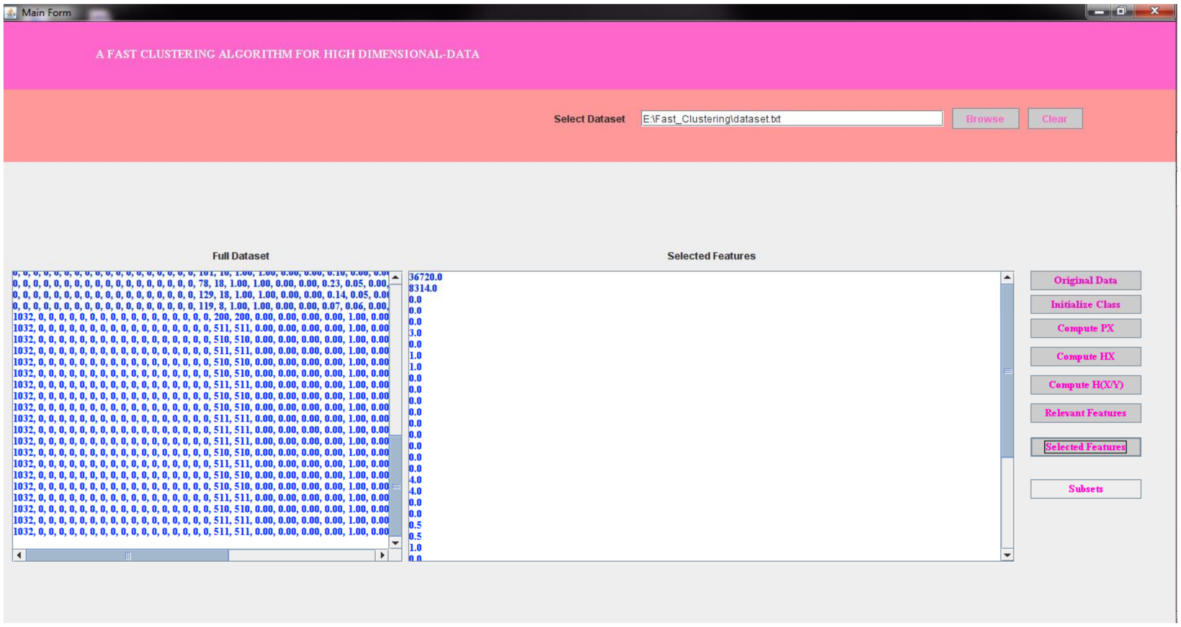
Relevant Features

```
0.0
0.0
511.0
511.0
0.0
0.0
0.0
0.0
1.0
0.0
0.0
255.0
113.0
0.44
0.09
0.44
0.0
0.0
0.0
0.0
0.0
0.05
1032.0
0.0
0.0
```

Figure 4.2 Selections of Relevant Features

In this Figure 4.2 shows the relevant features selected by the RELEVANT clustering algorithm. The Symmetric Uncertainty (SU) is calculated for each feature in the original data. The feature having SU value less than the predefined threshold is removed as the irrelevant feature from the dataset then the remaining features are selected as the relevant features.

4.3. Redundant Features Elimination



Select Dataset E:\Fast_Clustering\dataset.txt

Full Dataset

```
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 101, 10, 1.00, 1.00, 0.00, 0.00, 0.10, 0.00, 0.0
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 78, 18, 1.00, 1.00, 0.00, 0.00, 0.23, 0.05, 0.00,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 129, 18, 1.00, 1.00, 0.00, 0.00, 0.14, 0.05, 0.0
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 119, 8, 1.00, 1.00, 0.00, 0.00, 0.07, 0.06, 0.00,
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 200, 200, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 510, 510, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00
```

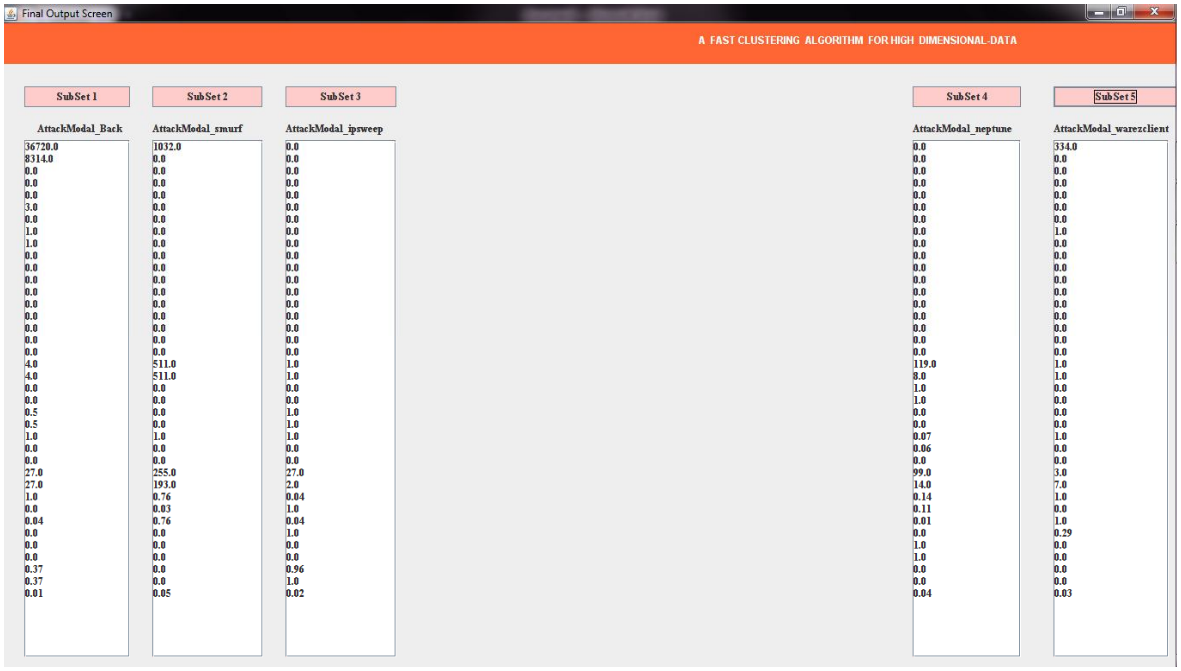
Selected Features

```
36720.0
8314.0
0.0
0.0
0.0
3.0
0.0
1.0
1.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
4.0
4.0
0.0
0.0
0.5
0.5
1.0
0.0
```

Figure 4.3: Redundant Features Elimination

In this Figure 4.3 shows the redundant features elimination. From the relevant features the redundant features are removed .The relevant features are used to construct the Minimum Spanning Tree (MST).Partitioning the MST to form the clusters.

4.4: Selection of features from the Clusters to form Feature Subsets



```
AttackModal_Back  AttackModal_smurf  AttackModal_ipsweep  AttackModal_neptune  AttackModal_warezclient
36720.0           1032.0             0.0                  0.0                  334.0
8314.0            0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
3.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
1.0               0.0                0.0                  0.0                  1.0
1.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
0.0               0.0                0.0                  0.0                  0.0
4.0               511.0              1.0                  119.0                1.0
4.0               511.0              1.0                  8.0                  1.0
0.0               0.0                0.0                  1.0                  0.0
0.0               0.0                0.0                  1.0                  0.0
0.5               0.0                1.0                  0.0                  0.0
0.5               0.0                1.0                  0.0                  0.0
1.0               1.0                1.0                  0.07                 1.0
0.0               0.0                0.0                  0.06                 0.0
0.0               0.0                0.0                  0.0                  0.0
27.0              255.0              27.0                 99.0                 3.0
27.0              193.0              2.0                  14.0                 7.0
1.0               0.76               0.04                 0.14                 1.0
0.0               0.03               1.0                  0.11                 0.0
0.04              0.76               0.04                 0.01                 1.0
0.0               0.0                1.0                  0.0                  0.29
0.0               0.0                0.0                  1.0                  0.0
0.0               0.0                0.0                  1.0                  0.0
0.37              0.0                0.96                 0.0                  0.0
0.37              0.0                1.0                  0.0                  0.0
0.01              0.05               0.02                 0.04                 0.03
```

Figure 4.4: Selection of features from the Clusters to form Feature Subsets

In this Figure 4.4 shows the five subsets consists of five different attacker features. The five attacker subsets are Back attack, smurf attack, ipsweep attack, Neptune attack and ware client attack. From the redundant features elimination features are selected from cluster to form the subsets of features.

V. CONCLUSIONS

In this paper, we have proposed a RELEVANT clustering algorithm for high dimensional data. The algorithm involves 1) selecting relevant features, 2) constructing a minimum spanning tree from relevant features, and 3) partitioning the MST and selecting features from cluster to form subset of features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and dissimilar to the objects belonging to other clusters thus dimensionality is drastically reduced.

The performance of the proposed RELEVANT clustering algorithm is compared with those of the well-known feature selection and clustering algorithms such as FCBF, Relief, CFS, INTERACT, FOCUS-2. The RELEVANT clustering algorithm in comparison provides the result that it is more efficient and effective than the feature selection algorithm for finding the subset of features.

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] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection,"Artificial Intelligence,vol. 97, nos. 1/2, pp. 273-324, 1997.
[12] D. Koller and M. Sahami, "Toward Optimal Feature Selection,"Proc. Int'l Conf. Machine Learning,pp. 284-292, 1996.
[13] I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF,"Proc. European Conf. Machine Learning,pp. 171-182, 1994.
[14] C. Krier, D. Francois, F. Rossi, and M. Verleysen, "FeatureClustering and Mutual Information for the Selection of Variables in Spectral Data," Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning,pp. 157-162,2007.
[15] L. Yu and H. Liu, "Efficiently Handling Feature Redundancy in High-Dimensional Data,"Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '03),pp. 685-690, 2003
[16] Z. Zhao and H. Liu, "Searching for Interacting Features,"Proc.20th Int'l Joint Conf. Artificial Intelligence,2007.