An Efficient Clustering based FAST Algorithm for High Dimensional Data

Anil S Naik
Department of Information Technology
Walchand institute of technology, Solapur, Maharashtra, India

Dayanand S Patil
Department of Computer Science Engineering
Walchand institute of technology, Solapur, Maharashtra, India

Abstract— Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, ReliefF, CFS with respect to four types of well-known classifiers, namely, the probability based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. The results, on 2 publicly available text data, demonstrate that the FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.

Keywords – Data Mining, Clustering, Feature selection Algorithm, ReliefF, CFS

I. INTRODUCTION

In this paper, we use a Efficient Clustering Method for High dimensional Data, based on feature selection identifies a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm can be evaluated with respective of the efficiency and effectiveness. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, an Efficient Clustering Method for High Dimensional Data is proposed and experimentally evaluated in this paper. This method works in two steps.

In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent the clustering-based strategy of Efficient Clustering Method for High dimensional Data has a high probability of producing a subset of useful and independent features.
In machine learning and statistics, 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. 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 (or data points). The archetypal case is the use of feature selection in analyzing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples.

Feature selection techniques provide three main benefits when constructing predictive models:

- Improved model interpretability.
- Shorter training times.
- Enhanced generalization by reducing overfitting.

Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction, and how these features are related. The past approach there are several algorithm which illustrates how to maintain the data into the database and how to retrieve it faster, but the problem here is no one cares about the database maintenance with ease manner and safe methodology.

Existing System: The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. 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. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. Disadvantages of the existing system are the generality of the selected features is limited and the computational complexity is large and the accuracy of the learning algorithms is not guaranteed.

II. PROPOSED ALGORITHM

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 and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features.
features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed efficient clustering algorithm for High dimensional data 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.

**Advantages:**
- 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.

**Algorithm Used:**
The Efficient clustering algorithm for High Dimensional Data.
The proposed algorithm logically consists of three steps:
- Removing irrelevant features and redundant features.
- Constructing a MST from relative ones and
- Partitioning the MST and selecting representative features.

In the first step it will identify data source from huge dataset in data warehouse and removes irrelevant features and redundant features from identified data source. In the second step it will construct minimum spanning tree for identified data, these tree forms a forest. And each MST forms cluster based target class features. From that it selects most accurate predicted details or data.

### III. SYSTEM ARCHITECTURE

The extraction of hidden predictive information from large database is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. 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. Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed Efficient Clustering algorithm for High-dimensional data falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features.

**Irrelevant feature removal:** In this module, first eliminate the irrelevant features in the given data set using this algorithm. Find out the relevance between each feature and the target concept. That means calculate the distance between each and every features and the target concept. The distance is calculated using Euclidian and Manhattan distance. If the distance is greater than the predetermined threshold value then it is a relevant feature. Otherwise it is a irrelevant feature. In this way we have to eliminate the irrelevant and get the relevant ones.

**Redundant feature removal:** In this module which composed of three components.
1. Minimum Spanning Tree Construction
2. Tree Partition or Clustering
3. Representative feature selection

**Minimum spanning tree construction:** In this module, construct the MST from the relevant features. After we eliminate the irrelevant features we get the relevant features. In this relevant features taking a pair of features. Apply the feature correlation between these features. Calculate the information gain using entropy technique. After calculate the information gain of the pair of features then choose high information gain feature other features are removed from the relevant ones. Finally we get highly correlated features with target concept. Construct the MST from these relevant features using prim algorithm.
Figure 2: System Architecture

Data flow diagram: A data flow diagram (DFD) is a graphical representation of the “flow” of data through an information system. It differs from the flowchart as it shows the data flow instead of the control flow of the program. A data flow diagram can also be used for the visualization of data processing. The DFD is designed to show how a system is divided into smaller portions and to highlight the flow of data between those parts.

Function

Input

Search Dataset

Output

Associated Output

Distributed Clustering

1. Subset Selection Algorithm
2. Time Complexity
3. Text Search.
4. Image search.(Ontology Based Search)

IV. FAST ALGORITHM AND ITS EXECUTION

- Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines
- 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
- In our proposed FAST algorithm, it involves (a) the construction of the minimum spanning tree (MST) from a weighted complete graph; (b) the partitioning of the MST into a forest with each tree representing a cluster; and (c) the selection of representative features from the clusters
Algorithm 1: FAST

| inputs: D(F₁, F₂, ..., Fₘ, C) - the given data set |
| θ - the T-Relevance threshold. |
| output: S - selected feature subset. |

/* === Part 1: Irrelevant Feature Removal ===*/
1 for i = 1 to m do
2    if T-Relevance = SU(Fᵢ, C) > θ then
3        S = S U {Fᵢ};

/* === Part 2: Minimum Spanning Tree Construction ===*/
5 G = NULL; //G is a complete graph
6 for each pair of features {Fᵢ', Fᵢ''} ∈ S do
7    F-Correlation = SU(Fᵢ', Fᵢ'')
8    Add Fᵢ' and/or Fᵢ'' to G with F-Correlation as the weight of the corresponding edge;

9 minSpanTree = Prims(G); //Using Prims Algorithm to generate the minimum spanning tree

/* === Part 3: Tree Partition and Representative Feature Selection ===*/
10 Forest = minSpanTree
11 for each edge Eᵢj ∈ Forest do
12    if SU(Fᵢ', Fᵢ'') < SU(Fᵢ', C) ∧ SU(Fᵢ'', C) < SU(Fᵢ', C) then
13        Forest = Forest - Eᵢj
14 S = φ
15 for each tree Tᵢ ∈ Forest do
16    Fᵢ̃ = argmax Fᵢ' ∈ Tᵢ SU(Fᵢ', C)
17    S = S U {Fᵢ̃};
18 return S

Figure 4: FAST algorithm
Figure 5: Original High dimensional data set of Heart before applying on FAST algorithm for execution.

Figure 6: High dimensional text data set of Heart execution using FAST algorithm.
V. RESULTS AND ANALYSIS

In this section we present the experimental results in terms of the proportion of selected features (data set), the time to obtain the feature subset. Fast algorithm is more efficient than CFS and ReliefF Algorithm.
VI. CONCLUSION AND FUTURE ENHANCEMENT

In this paper presented a novel Efficient clustering method for high dimensional data method involves removing irrelevant features, Constructing a minimum spanning tree from relative ones, and Partitioning the MST and Selecting representative features. In this method a cluster consists of features, each cluster is treated as a single feature and thus dimensionality is drastically reduced. By choosing a subset of good features with respect to the target concepts, and feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility, Efficient Clustering method for high dimensional data is very effective and efficient.

For the future work, we can explore different types of correlation measures, and study some formal properties of feature space. In feature we are going to classify the high dimensional data.

REFERENCES