RESEARCH ARTICLE OPEN ACCESS

A Survey on Constellation Based Attribute Selection Method for High Dimensional Data

Ms.Sonam R. Yadav*, Prof. Ravi Patki**
*(Department of Computer Engineering, Savitribai Phule Pune University, Pune)
** (Department of Information Technology, Savitribai Phule Pune University, Pune)

ABSTRACT
Attribute Selection is an important topic in Data Mining, because it is the effective way for reducing dimensionality, removing irrelevant data, removing redundant data, & increasing accuracy of the data. It is the process of identifying a subset of the most useful attributes that produces compatible results as the original entire set of attribute. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar in some sense or another to each other than to those in other groups (Clusters).

There are various approaches & techniques for attribute subset selection namely Wrapper approach, Filter Approach, Relief Algorithm, Distributional clustering etc. But each of one having some disadvantages like unable to handle large volumes of data, computational complexity, accuracy is not guaranteed, difficult to evaluate and redundancy detection etc.

To get the upper hand on some of these issues in attribute selection this paper proposes a technique that aims to design an effective clustering based attribute selection method for high dimensional data. Initially, attributes are divided into clusters by using graph-based clustering method like minimum spanning tree (MST). In the second step, the most representative attribute that is strongly related to target classes is selected from each cluster to form a subset of attributes. The purpose is to increase the level of accuracy, reduce dimensionality; shorter training time and improves generalization by reducing over fitting.

Keywords – Attribute Selection, Clustering, Data Mining, Graph-based Clustering, Minimum Spanning Tree.

I. INTRODUCTION

Attribute selection is the process of identifying a subset of the most useful features that produces compatible results as the original entire set of features. An attribute 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 attributes, the effectiveness is related to the quality of the subset of attributes [1]. It is an important topic in Data Mining, because it is the effective way for reducing dimensionality, removing irrelevant data, removing redundant data, & increasing accuracy of the data.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other clusters. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [2].

Attribute selection, also known as feature selection or variable subset selection. The process of selecting a subset of relevant attributes is used in model construction. The assumption when using an attribute selection technique is that the data contains many redundant or irrelevant (noisy) attributes. Redundant attributes and irrelevant attributes provide no useful information in any context and it diminishes the quality of the selected attributes [5]. Attribute selection techniques are a subset of the more general field of attribute extraction in high dimensional data. Attribute extraction creates new data from original attributes of dataset, whereas attribute selection returns a subset of the attributes. Attribute selection techniques are frequently used in domains where there are many attributes and comparatively few samples.

An attribute selection algorithm proposed in this paper can be seen as the combination of a search technique for proposing new feature subsets along with an efficiency and effectiveness. The simplest step is to test each possible subset of attributes finding the one which increases the accuracy.

The proposed algorithm consists of three parts:

(i) removing irrelevant attributes
(ii) constructing a MST from relative one
(iii) Partitioning the MST and selecting representative attributes.

II. BASIC CONCEPTS

In this section, a brief introduction about the basic concepts of the Data Mining, Clustering, and Minimum Spanning tree is provided.

2.1 Data Mining:

Data mining is the way of discovering the interesting knowledge from large amounts of information sources or data warehouses. When there is huge amount of data and certain information is to be found out from that data, then different stages of data mining is applied to it to gain information from it. Data mining tasks classified into two forms: 1. Descriptive mining tasks: Represent the general properties of the data. 2. Predictive mining tasks: Perform the implication on the current data.

Different Data mining Functionalities are: Characterization and Discrimination, Mining Frequent Patterns, Association and Correlations, Classification and Prediction, Cluster Analysis, Outlier Analysis, Evolution Analysis. Out of these functionalities this paper focuses on the cluster Analysis.

2.2 Cluster Analysis:

Clustering is the grouping similar objects into one class. A cluster is an association of data objects that are similar to one another within the same cluster and are dissimilar to the objects in different clusters. Document clustering (Text clustering) is closely related to the concept of data clustering. Document clustering is a more exact technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. Clustering helps to reduce the dimension and it simplifies the task as number of dataset is minimized to form a cluster.

2.3 Minimum Spanning Tree:

A minimum spanning tree (MST) is an undirected, connected, acyclic weighted graph with minimum weight. The idea is to start with an empty graph and try to add edges one at a time, the resulting graph is a subset of some minimum Spanning tree. Each graph has several spanning trees. This method is mainly used to make the appropriate attribute subset clustering but it take time to construct the cluster.

Various Applications:

- Design of computer networks and Telecommunications networks
- Transportation networks, water supply networks, and electrical grids.
- Cluster analysis
- Constructing trees for broadcasting in computer networks
- Image registration and segmentation

III. BACKGROUND AND COMPARATIVE ANALYSIS

Many algorithms & techniques have been proposed up till now for clustering based feature/ attribute selection. Some of them have focused on minimizing redundant data set and to improve the accuracy whereas some other features subset selection algorithm focuses on searching for relevant features

Fuzzy Logic used for improving the accuracy. In this, after removal of redundant data, clustering is done based on Prim's Algorithm. It results in MST (Clusters) which guarantees redundancy. Once we get the minimized redundant data sets, accuracy automatically gets increases [3].

An algorithm for filtering information based on the Pearson test approach has been implemented and tested on feature selection. This test is frequently used in biomedical data analysis and should be used only for nominal (discretized) features. FCBF (Fast Correlation Based Feature Selection) algorithm is the modified algorithm, FCBF#, has a different search strategy than the original FCBF and it can produce more accurate classifiers for the size *k* subset selection problem [4].

Relief is well known and good feature set estimator. Feature set estimators evaluate features individually. The fundamental idea of Relief algorithm is estimate the quality of subset of features by comparing the nearest features with the selected features. With nearest hit (H) from the same class and nearest miss (M) from the different class perform the evaluation function to estimate the quality of features. This method used to guiding the search process as well as selecting the most appropriate feature set.

Hierarchical clustering is a procedure of grouping data objects into a tree of clusters. It has two types: 1) Agglomerative approach is a bottom up approach; the clustering processes starts with each object forming a separate group and then merge these atomic group into larger clusters or group, until all the objects are in a single cluster. 2) Divisive approach is reverse process of agglomerative; it is top down approach, starts with all of objects in the same cluster. In each iteration, a clusters split up into smaller clusters. And finally from the individual clusters subset is taken out.

Affinity Propagation Algorithm which is used to solve wide range of clustering problems. Algorithm proceeds in the same way i.e. removes irrelevant data, then construct minimum spanning tree and partition the same and select the most effective features. This algorithm aims to reduce the dimensions of the data. In this, data is clustered depending upon their similarities with each other using pair wise constrain. A cluster consists of

features. And then each cluster is treated as a single feature and thus dimensionality is reduced.

Table 3.1 Comparison of Previous Techniques

S.NO	**Techniques (or)Algorithms**	**Advantages**	**Disadvantages**
1.	Consistency Measure	Fast, Remove noisy and irrelevant data	Unable to handle large volumes of data
2.	Wrapper Approach	Accuracy is high	Computational complexity is large
3.	Filter Approach	Suitable for very large features	Accuracy is not guaranteed
4.	Agglomerative linkage algorithm	Reduce Complexity	Decrease the Quality when dimensionality become high
5.	INTERACT Algorithm	Improve Accuracy	Only deal with irrelevant data
6.	Distributional clustering	Higher classification accuracy	Difficult to evaluation
7.	Relief Algorithm	Improve efficiency and Reduce Cost	Powerless to detect Redundant features

Comparison of various algorithms and techniques are discussed as follows. Some of them are having low computational complexity, higher accuracy or some are having redundancy problem while some are good at classification level but difficult to evaluate.

IV. PROPOSED SYSTEM

The proposed clustering-based attribute selection algorithm is based on minimum spanning tree (MST). The proposed algorithm works in two steps. In the first step, attributes are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative attribute that is strongly related to target classes is selected from each cluster to form a subset of attributes. Attributes in different clusters are relatively independent; the clustering-based strategy of proposed algorithm has a high probability of producing a subset of useful and independent attributes.

Proposed algorithm is a straight-forward method in which main basic concept used is clustering. The need for clustering is to find the closest attribute selection among large volume of data i.e. over the high dimensional data. In business analysis, in hospitals for patient record maintenance, in banking system and for various marketing purposes high dimensional data is been used. Extraction of one particular file related to specific thing or person, from such huge amount of high dimensional data is very difficult and time consuming. High dimensional data refer to multiple attributes to each of the record i.e. suppose in a banking system each account holder is having various attributes related to account. For example Account details including Account number, account type, customer number, amount, last date of transactions, last time of transaction, type of transaction etc. So, from such huge amount of data retrieval is bit complex and time required is more. And in applications like banking system, patients record maintenance; time is the most important factor. And another important factor is accuracy.

Proposed algorithm can efficiently and effectively deal with both irrelevant and redundant attributes, and obtain a good attribute subset. The irrelevant attribute removal is straightforward once the right relevance measure is defined or selected, while the redundant attribute elimination is a bit of sophisticated.

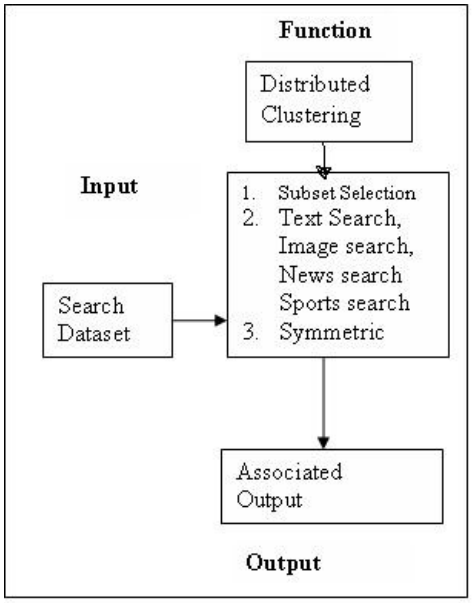

Figure 4.1 System Architecture

In proposed algorithm, it involves

1. The construction of the minimum spanning tree from a weighted complete graph;
2. The partitioning of the MST into a forest with each tree representing a cluster; and
3. The selection of representative attributes from the clusters.

System architecture consists of different phases of the processing. Overall procedure includes construction of minimum spanning tree followed by partitioning of the same and finally the process of selection of attributes from the clusters.

V. UTILITY

- Proposed system may be used across all domains over unlimited data set to search for the effective and best result.
- Proposed system would be very useful in various hospitals for patient record maintenance which is nothing but high dimensional data and from that irrelevant and redundant features have to be removed and finally selected features have to be produced.
- The same is applicable in banking system where the data is high dimensional and for each transaction or process, access time and accuracy must be very high.
- Proposed system may be used to for various marketing purposes where there is a lot of data and the only relevant and suitable data have to be published each time.
- Proposed system is useful when there is high dimensional data, and a pattern or a result is to be finding out from that large amount of data in less time with greater accuracy.

VI. CONCLUSION

Proposed algorithm describes a systematic workflow for the attribute selection over large amount of data sets. It uses minimum spanning tree to do so. In this paper, we have presented a novel clustering-based attribute subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant attributes, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative attributes. Each cluster is treated as a single attribute and thus dimensionality is drastically reduced.

REFERENCES

[1] N.Magendiran and J.Jayaranjani, An Efficient Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data - (*ICETS'14*)
[2] Mr. M. Senthil Kumar and Ms. V. Latha Jothi, A Fast Clustering Based Feature Subset Selection Using Affinity Propagation Algorithm - (ICGICT'14)

[3] T.Jaga Priya Vathana, C. Saravanabhavan, and Dr.J. Vellingiri, A Survey On Feature Selection Algorithm For high Dimensional Data Using Fuzzy Logic - (IJES)
[4] Lei Yu and Huan Liu,Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
[5] S.Swetha and A.Harpika, A Novel Feature Subset Algorithm For High Dimensional Data – (IJRRECS)
[6] 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, 2004s
[7] 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
[8] Saurabh Soni & Pratik Patel, "IFSS – An Improved Filter-Wrapper Algorithm for Feature Subset Selection", International Journal of Computer Application (0975-8887), Volume 95-No. 14, June 2014.