ISSN:0975-9646

Efficient Data Storing and Retrieving from University Database using Feature Selection and Clustering Method

Vaidehi Ambekar, Prof. P.S.Mohod

Department of Computer Science and Engineering, RTMNU

***Abstract*—— Unstructured Data refers to information that neither have a pre-defined data model nor is organized in a pre-defined manner. These results in irregularities and unclarified that does not make it so easy to understand using traditional computer programs as comparing with data stored in fielded form in databases or annotated in documents. Hence Feature selection technique is applied on this unstructured data. Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. It is very hard to process over the large data set with respect to the time, memory and the computation cost. Because the bigger size data set are having the large number of the attributes to process, it is not necessary in many cases that all the features are relevant to get the knowledge about the domain. Feature selection identifies the subset of data that are relevant to parameter called Maximum Relevance. Such subset frequently containing material as it is relevant but redundant so it attempts to mark this problem by eliminating those redundant subsets. The purpose of clustering is to reveal the natural structure inherent data and extracting useful information from noisy data.**

Index Terms— Feature selection, graph-based clustering

I. INTRODUCTION

It is very hard to process over the large data set with respect to the time, memory and the computation cost because the bigger size data set are having the large number of the attributes to process, it is not necessary in many cases that all the features are relevant to get the knowledge about the domain. We want to decrease the number of features to process. Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. Feature selection, the process of removing irrelevant features can be extremely useful in reducing the dimensionality of the data to be processed, in reducing execution time and improving predictive accuracy of the classifier. Unstructured Data refers to information that neither have a pre-defined data model nor is organized in a pre-defined manner. that To understand using traditional computer programs as compared to data that are stored in fielded form in databases or annotated in documents do not make it so easy by these results in irregularities and unclarified. Hence on unstructured data this Feature selection technique is applied. To extract useful information from noisy data and showing the natural structure of inherent data is the main purpose of clustering.

2. LITERATURE REVIEW

For generating many small size clusters with high term association as small real index size which is able to store into different nodes as a whole so index clustering algorithm is in designed in paper[1]. To map different logical index clusters onto physical nodes a practical index mapping algorithm is proposed for keeping high query locality with reasonable load balancing. For large scale astronomical data index system stimulation results shows algorithm provided in this paper having good scalability.

In paper [2], Highlighting the activities done by bioinformatics community in a developing novel with adapting procedure reviews the most important application in bioinformatics domain to different feature selection techniques for classification. Biological knowledge implies the statistical significance of enrichment of GO terms in the cluster as represented by the constitute gene .The algorithms CLARANS were employed as a tool for attribute clustering in large gene space. Extraction of subset of genes, from high-dimensional gene space, lead to reduced computational complexity .It is to be noted that computationally efficient globally optimum measure of evaluation does not necessary converge to a biological meaningful solutions.

In paper [3],The FAST algorithm execute in two steps is proposed and experimentally evaluates by fast clustering-based feature selection algorithm(FAST). In the second step, cluster formation is done by selecting the most standard feature that is forcibly related to target classes. While in first step by using graph-theoretic clustering method divides features into clusters. Features in different clusters are not dependent, producing a subset of useful and not dependent features have high probability with the clustering-based strategy of FAST. FAST algorithm efficiency is improved by minimum spanning tree method(MST).Publicly available in real-world high-extent image, text data and microarray, demonstrate that FAST improves the performances of the four types of classifiers and produces smaller subsets of features .For the future work, some formal properties of feature space and explore different types of correlation measures.

3.FRAMEWORK AND DEFINITIONS:

Feature Selection:

Accuracy of the learning machines severely affects the irrelevant features, along with redundant features. Thus, feature subset selection can identify and remove as much of the redundant and irrelevant information as possible. Moreover, “good attribute subsets contain attribute

uncorrelated with (not predictive of) each other and highly correlated with (predictive of) the class." We develop a novel algorithm, which can effectively and efficiently deal with both redundant and unrelated features, and obtain relevant feature subset. We achieve this through a new feature selection framework (shown in Fig. 1) which is made up of the two connected components of unwanted feature removal and redundant feature elimination. Features obtained by the former relevant to the target concept by elimination of irrelevant ones, and different feature cluster is represented by relative one, thus produces the final subset the latter removes redundant features from relevant ones.

The irrelevant feature removal is straightforward, while the redundant feature elimination is a bit of sophisticated once the right relevance measure is defined or selected. In our proposed FAST algorithm, it involves -

1) MST is partitioned into a forest where cluster is represented by each tree; and
2) Weighted complete graph is form by construction of minimum spanning tree;
3) cluster is form by selection of representative feature.

Our proposed feature subset selection framework involves redundant feature elimination and irrelevant feature removal for more precisely introduction of the algorithm. In this paper, a new validity index, which determines the number of clusters for semantic hierarchical clustering is proposed. The method is applied for automatic categorization. Our focus is on strings in a single word, based on which processing on strings in multiple words is also applicable. Since words lack of content for statistical conclusion, we employ WordNet to get se-mantic similarity directly. The main contribution of this paper is to introduce a new validity index in keyword clustering.

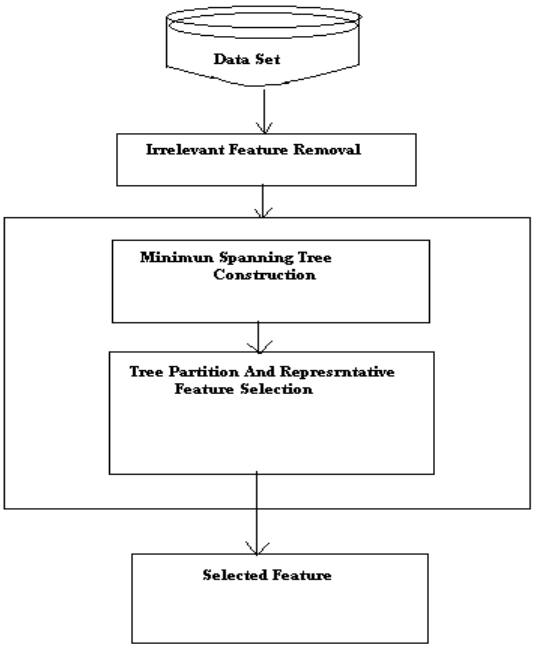

Fig. 1.Framework of the proposed feature subset selection algorithm

Maximum Relevance Minimum Redundancy:

Feature selection, machine learning and pattern recognition is one of the basic problems which identifies data that are appropriate to the parameters used and is normally called *Maximum Relevance*. In this project Maximum relevance is refers to the data in well organized form, redundant data refers to the unorganized form. By removing those redundant subsets these subsets often contain material which is relevant but redundant. In many areas such as speech recognition and cancer diagnosis mRmR has a variety of applications.. These subsets often contain data which is redundant but relevant and mRmR to address this problem it removes redundant feature.

4.CLUSTERING ALGORITHMS:

In many fields clustering is a common unsupervised learning method. Into two different types clustering methods are traditionally classified: hierarchical clustering and partition clustering. Hierarchical tree is the output of agglomerative hierarchical clustering algorithm. The algorithm needs to compute the equale scores of all pairs of data items to cluster the data items and form a hierarchical tree, which are stored in a similarity matrix for future reference. Then, the algorithm merges the data items from the most similar pair to the most dissimilar pair to grow a hierarchical tree. Computation of the similarity scores requires a time complexity of O(n2), and typically, the similarity scores have to be stored for future references, implying that the space complexity is O(n2) as well. For example, a data set with 300, 000 data items need at least 360 gigabytes if the similarity scores are saved in float. For many scientific and engineering applications that only consider highly related data, only a small portion of the most similar pairs of data items is referenced in clustering. Following this idea, we propose a parallel disjoint set approach to compute the similarity matrix in parallel, and use a threshold to prune unrelated data items to reduce the required space for the similarity matrix. The proposed clustering algorithm then uses this reduced similarity matrix to sequentially create disjoined sets of highly related data items. Each disjoined set is clustered in parallel to form a hierarchical sub-tree. Finally, similar score is compute by algorithm to form a hierarchical tree. Through the similarity matrix and the disjoint set clustering, we find all the disjoint sets under the threshold. To construct a hierarchical tree to represent the similarity relationship among the disjoint sets, two tasks have to be accomplished. First, hierarchical tree is formed by disjoint sets.

Second, the merging of sub-tree is taking place according to the similarity. After all possible disjoint sets are clustered into many sub-trees, another similarity matrix of these sub-trees is constructed. Then a complete hierarchical tree is constructed according to the new similarity matrix of the sub-trees.

Graph-theoretic clustering methods:

In the case of community college students, for example, clusters may be formed on the basis of student aspirations ,student course-taking and enrolment behaviours , student demographic characteristics , or any

combination of these or other measures. With respect to the feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. 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. The result is a forest and each tree in the forest represents a cluster. Hierarchical constrained clustering of spectral variables and selection of clusters by mutual Information. Their feature clustering method is similar to every cluster to contain consecutive featuresonly.Both methods employed agglomerative hierarchical clustering to remove redundant features.

5. LATENT SEMANTIC INDEXING:

Principle of LSI is that it uses same text for similar meaning. Relationship between similar texts is achieved by extracting similar content of body text. LSI has capability to find close semantic similarity between related terms that are latent in a collection of text uncovers the underlying latent semantic structure in the usage of words in a body of text and how it can be used to extract the meaning of the text in response to user queries, commonly referred to as concept searches. Set of document such as queries, or concept searches, that go through LSI will give a result which is having are conceptually similar in meaning with search text. It is not necessary that result should share a specific words with query entered by user.LSI deal with Boolean keyword queries: multiple words having similar meanings, words having more than one meaning. Automated document categorization is performed by LSI. Document categorization is classifying the document to one or more predefined categories base on similarity and conceptual content of the categories. During categorization processing, documents having correlated terms are categorized, then this compared to the concepts contained in the example items, similar concepts documents assigned to the same categories.

6.CONCLUSION:

In this paper, we have presented a novel clustering and feature subset selection algorithm for huge amount data. The algorithm involves 1) removing irrelevant attribute, 2) From relative attribute construct minimum spanning tree3)selection representative feature by dividing the MST. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced

REFERENCES:

1] Hsi-Che Liu, Pei-Chen Peng,"Comparison of feature selection methods for cross-laboratory microarray analysis" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS,2013

2] Sushmita Mitra, ,"Feature Selection and Clusterind of Gene Expression Profile UsingBiological Knowledge" IEEE TRANSACTIONS ON SYSTEM ,MAN ,AND CYBERNETICS.NOVEMBER 2012

[3] Qinbao Song, Jingjie Ni, and Guangtao Wang "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013

4] T.Chandrasekhar "Unsupervised Gene Expression Data using Enhanced Clustering Method" IEEE International Conference on Emerging Trends in Computing, 2013

[5] Cosmin Lazar"Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 9, NO. 4, JULY/AUGUST 2012

[6] Cheng-Hsien Tang, An-Ching Huang"An Efficient Distributed Hierarchical-Clustering Algorithm for Large Scale Data" 2010 IEEE