SURVEY ON HIGH DIMENSIONAL DATA CLUSTERING USING FAST CLUSTER BASED FEATURE SELECTION

Bhuvaneshwari Melinamath[1], Shruti Hebbal[2]

[1]Assistant Professor, [2]M.Tech Scholar, Department of Computer Science & Engineering,
BLDEA's Dr.P.G.Halakatti College of Engineering and Technology, Vijayapura, Karnataka (India)

ABSTRACT

High-dimensional data often contain irrelevant or redundant features which slow down the mining process and cause difficulties in storage and retrieval. Feature selection is the process of selecting most relevant features from an entire set of features. The FAST (fast clustering –based feature selection algorithm) 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 subset from each cluster to form a subset of features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST using the kruskal's algorithm) clustering method.

I. INTRODUCTION

With the rapid growth of computational biology and e-commerce application, high-dimensional data becomes very common. The major challenge of high dimensional data is its curse of dimensionality. The complexity of many existing data mining algorithm is exponential with respect to the number of dimensions.

In the literature many approaches have been proposed for dimensionality reduction. The existing dimensionality reduction methods can roughly be categorized into two classes: feature extraction and feature selection. In feature extraction problems, the original features in the measurement space are initially transformed into a new dimension-reduced space. Although the significant variables are related to the original variables, the physical interpretation in terms of the original variables may be lost.

Feature selection aims to seek optimal or suboptimal subsets of the original features, by preserving the main information carried by the collected complete data, to facilitate future analysis for high dimensional problems.

Feature selection involves searching through various feature subsets, followed by the evaluation of each of them using some evaluation criteria. The mostly used search strategies are greedy sequential searches through the feature space, either forwards or backwards. Different types of heuristics, such as sequential forward or backward search, floating search, beam search, bidirectional search, and genetic search, have been suggested to navigate the possible feature subsets .In supervised learning ,classification accuracy is widely used as evaluation criterion. However, in unsupervised learning feature selecting is more challenging since the class labels are unavailable to guide the search.

Feature selection algorithms can be broadly classified into the filter model and the wrapper model. The filter model and the wrapper model .The filter model rely on general characteristics of the training data to select some

features without involving any learning algorithm. The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected.

The ability to quickly and effectively process large amounts of data is necessary in order to effectively scale learning algorithm to match the growth of data available. Clustering algorithms are an unsupervised machine learning technique that facilitates the creation of clusters, which allow us to group similar items together so that these. Clusters are similar in some sense. Clustering has broad applications in areas such as data mining, recommendation systems pattern recognition, identification of abnormal cell clusters for cancer detection, and bioinformatics.

1.1 System Architecture

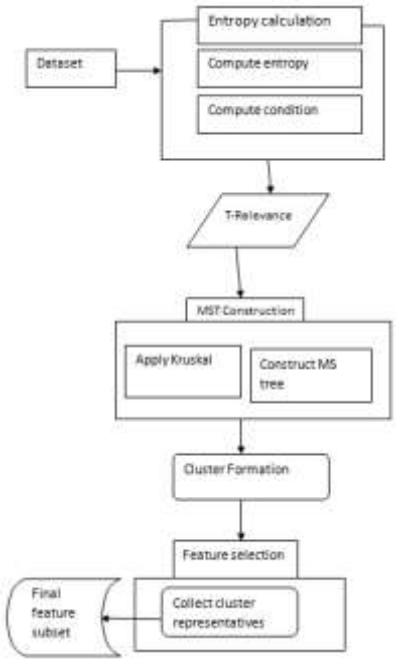

Fig 1. FAST Frameworks [1]

To remove irrelevant feature and redundant feature, the FAST algorithm has two connected components. Irrelevant feature removal and redundant feature elimination. The irrelevant feature removal is straight forward once the right relevance measure is defined or selection while the redundant feature elimination is a bit of sophisticated. In this FAST algorithm, it involves

- The construction of the minimum spanning tree from a weighted complete graph
- The partitioning of the MST into a forest with each representing a cluster and
- The selection of representative features from the clusters
- Load Data

The data has to be pre-processed for removing missing values, noise and outliers. Then the given dataset must be converted into the arff format. From the arff format, only the attributer and the values are extracted and stored into the database.

- Entropy and Conditional Entropy Calculation

Relevant feature have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation. To find the relevance of each attribute with class label, Information gain is computed. This is also said to be mutual information (MI) measure.

MI measure how much the distribution of the feature values and target classes differ from statistical independence. The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values and target classes, and has been used to evaluate the goodness of feature for classification. The SU is defined as follows:

$$SU(X,Y)=2*Gain(X|Y)H(X)+H(Y)$$

Where, H(X) is the entropy of a random variable X. Gain (X|Y) is the amount by which the entropy of Y decreases.

- T-Relevance and F-Correlation computation

The relevance between the feature Fi and the target concept C is referenced to as the T-Relevance of Fi and C, and denoted by SU (Fi, C). If SU (Fi, C) is greater than a predetermined threshold, then Fi is a strong.

- Minimum spanning tree(MST) Construction

With the F-correlation value computed above, the MST is constructed. A MST is a sub-graph of a weighted, connected and undirected graph. It is acyclic, connect all the nodes in the graph, and the sum of all of the weight of all of its edges is minimum. That is, there is no other spanning tree, or sub-graph which connects all the nodes and has a smaller sum. If the weights of all the edges are unique, then the MST is unique. The nodes represent the samples, and the axis of the n-dimensional graph represents the n features. The complete graph G reflects the correlations among all the target-relevant features. MST is constructed using well-known Kruskal's algorithm.

II. RELATED WORK

Kathleen Ericson(2012) [1]:This paper provides experimental setup for clustering based on hadoop (version 1.0.0) and Granules implementation and the data initially read from HDFS cluster. All tests are run on quad-core machines which is of 2.4GHz Granules support computations that will be executed successive rounds and while retaining state it is well suited for clustering algorithm that are iterative. Mahout Naive and complementary bayes are implemented with distributed classification of algorithms which helps to determine the effect of moving a file based to a streaming based framework and these algorithms operates quickly and provide accurate recommendation in timely manner.

Anil K.Jain(2009) [6]: To have information about data clustering machine learning and pattern recognition communication are very important. For finding NP-hard problem we use the K-means algorithm which provides computationally efficient solution.

Yun zheng and Chee keong kwoh(2011) [5]: The paper mainly concentrated on the problem that helps in induction algorithms that are suffering from curse of dimensionality the redundancy and noisy attributes also results in lower performance and increased computation. To overcome curse of dimensionality feature selection algorithm is used.

S.Senthamarai Kannan (2007) [7]: Based on memetic framework a novel hybrid feature selection algorithm is proposed. Here filter ranking method is used as a local search heuristics. According to this paper the filter ranking method is a better approach then the Genetic algorithm (GA) and Memetic algorithm (MA).

Karthikeyan.P,saravanan.P,Vanitha.E(2014) [8]: In proposed FAST clustering based feature subset selection algorithm a cluster consist of features and each cluster is treated as single feature and hence dimensionality is drastically reduced. According to the experimental results feature subset selection algorithm not only reduces the number of features but also improves classification accuracy.

Yanhong Li,Ming Dong,Jing Hua (2007) [3]: Proposed feature selection algorithm which is relevant to all clusters but sometimes it is not applicable for many high dimensional datasets. This algorithm also provides better understanding of processes that generates the data. In this paper we made use of cross-projection method to have better quality of a individual clusters. According to the experimental results we cannot conclude that the clustering quality of LFS (localized feature selection) is worse than that of GFS (global feature selection) on this dataset as the error rate of LFS and GFS nearly similar.

GuangtaoWang,Baowen,BaowenXu,Yumi-ng Zhouu (2012) [9]: Proposed FOIL rule based feature subset selection algorithm which is applicable to high dimensional data. This algorithm is used not only for removing irrelevant and redundant feature but also with interactive features. The experimental results of the real world datasets show that our proposed algorithm has moderate reduction capability and it is much faster than that of other feature selection algorithm mainly on high dimensional data.

Hua-Liang Wei and Stephen A.Billings(2007) [10]: In this proposed new unsupervised learning algorithm for feature selection and dimensionality reduction. The main advantage of this algorithm is that implementation only involves the calculation of the designed correlation matrix and the forward orthogonalization procedure. It combines good effectiveness with high efficiency, often produces efficient feature subsets and thus, provides an effective solution to the dimensionality reduction.

III. CONCLUSION

The feature selection is a complex problem studied by many researchers all over the world. Complexity is due to finding a voluminous amounts of High Dimensional data, contains irrelevant or redundant features which causes difficulties in storage and retrieval. The feature subset selection algorithm for high dimensional data works based on the clusters that contains features where each cluster treated as single feature and hence dimensionality of data is drastically reduces.

REFERENCES

[1] ChinnuJose,M.Govindaraj and Cinu Skaria, ”An Enhanced Feature Selection Algorithm for High Dimensional data”, International conference on simulations in computing Nexus(ICSCN),(2014).

[2] Hua-Liang Wei and Stephen A.Billings, “Feature Subset Selection and Ranking for Data Dimensionality Reduction”, IEEE Transactions on pattern analysis and machine intelligence, vol.29, No.1, Jan-2007.

[3] Yuanhong Li,Ming Dong,JingHua, “Localized feature selection for clustering”,Elsevier-2007.

[4] KathleenEricson,ShrideepPallickara, ”On the performance of high dimensional data clustering and classification algorithm,Elsevier-2012.

[5] Yun Zheng and Chee Keong Kwoh, ”A Feature Subset Selection method Based on High-Dimensional Mutual information”, ISSN-2011.

[6] Anilk K.Jain,” Data Clustering: 50 years beyond K-means”, Elsevier-2009.

[7] S.Senthamarai Kannan,N.Ramaraj,” A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm”,Elsevier-2010.

[8] Karthikeyan.P,Saravanan.P,Vanitha.E, "High Dimensional Data Clustering using FAST Cluster Based feature selection, Journal of engineering research and application Vol.4,Mar-2014,pp.65-71.

[9] Guangtao Wang,Qinbao Song,Baowen Xu,Yuming Zhou, "Selecting feature subset for high dimensional data via the propositional FOIL rules"-Elsevier-2012.

[10] Hua-Liang wei and Stephen A.Billings, "Feature subset selection and Ranking for Data Dimensionality Reduction", IEEE Transactions on pattern analysis and machine intelligence, Vol. 29,N0.1 Jan-2007