S.Nagendrudu et al | IJCSET(www.ijcset.net) | May 2015 | Vol 5, Issue 5,113-115

ISSN : 2231-0711

Enhanced Clustering of High Dimensional Data Using Fast Cluster Based Feature Selection

S.Nagendrudu ,V.Ramakrishna Reddy

Computer Science Department, Santhiram Engineering College,JNTUA,,A.P., India

Abstract-A database can contain several dimensions or attributes. Many Clustering methods are designed for clustering low–dimensional data. In high dimensional space finding clusters of data objects is challenging due to the curse of dimensionality. When the dimensionality increases, data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered. To deal with these problems, an efficient feature subset selection technique for high dimensional data has been proposed.

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. The Minimum-Spanning Tree (MST) using Prim's algorithm can concentrate on one tree at a time. To ensure the efficiency of FAST, adopt the efficient MST using the Kruskal's Algorithm clustering method.

Keywords: Feature subset selection, filter method,feature clustering, graph-based clustering, Kruskal's algorithm

I.INTRODUCTION

Feature selection is an important topic in data mining, especially for high dimensional datasets. Feature selection (also known as subset selection) is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy. Feature selection [1] can be divided into four types: the Embedded, Wrapper, Filter, Hybrid approaches.

The Embedded methods incorporate feature selection as part of the training process and are usually specific to given learning algorithms. Decision Trees is the one example for embedded approach. Wrapper model approach uses the method of classification itself to measure the importance of features set, hence the feature selected depends on the classifier model used. Wrapper methods generally result in better performance than filter methods because the feature selection process is optimized for the classification algorithm to be used. However, wrapper methods are too expensive for large dimensional database in terms of computational complexity and time since each feature set considered must be evaluated with the classifier algorithm used [11]. The filter approach actually precedes the actual classification process. The filter approach is independent of the learning algorithm, computationally simple fast and scalable [11].

Filter strategies use a proxy live rather than the error rate to attain a feature set. This live is chosen to be quick to cypher, while still capturing the quality of the feature set. Common measures include the Mutual Information, Pearson product-moment coefficient of correlation, and inter/intra category distance[14]. Filters are sometimes less computationally intensive than wrappers, however they manufacture a feature set that isn't tuned to a selected style of prognosticative model. several filters give a feature ranking instead of an exact best feature set, and also the cutoff purpose within the ranking is chosen via crossvalidation

The Fast clustering Based Feature Selection algorithm (FAST) works into 2 steps. In the first step, features are divided into clusters by using graph theoretic clustering methods. In the second step, the most representative feature i.e., strongly related to target class is selected from each cluster to form the final subset of features. A feature in different clusters in relatively independent, the clustering based strategy of FAST has high probability of producing a subset of useful and independent features. The experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of four well known different types of classifiers. A good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. Different from these algorithms, our proposed FAST algorithm employs clustering based method to features.

This paper discusses about feature selection, FAST algorithm, Text classification and so on. The next section is literature review

II. LITERATURE REVIEW

A. An efficient approach to clustering in large multimedia databases with noise.

Several clustering algorithms can be applied to clustering in large multimedia databases. The effectiveness and efficiency of the existing algorithms, however is somewhat limited since clustering in multimedia databases requires clustering high-dimensional feature vectors and since multimedia databases often contains large amounts of noise. Using DENCLUE algorithm we can remove noise from the large databases. The main advantage includes, that the clustering can be done efficiently in high dimensional database.

B. Feature subset selection using the wrapper method: over fitting and dynamic search space topology

In the feature subset selection, a search for an optimal set of features is made using the induction algorithm as a black box. The best-first search is to find good feature subsets. The over fitting problems can be reduced by using the best

first search. The relevant and optimal features can be easily selected in this method, and also the over fitting can be reduced.

C. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining

Partitioning an outsized set of objects into unvaried clusters may be a basic operation in data processing. The k-means formula is best fitted to implementing this operation thanks to its potency in clump giant information sets. clump ways partition a collection of objects into clusters specified objects within the same cluster are a lot of almost like one another than objects in numerous clusters. the foremost distinct characteristic of information mining is that it deals with terribly giant data sets (gigabytes or maybe terabytes). This needs the algorithms utilized in data processing to be scalable.

D. Fast and Effective Text Mining Using Linear-time Document Clustering

Clustering is a powerful technique for large-scale topic discovery from text. It involves two phases: first, feature extraction maps each document or record to a point in highdimensional space, then clustering algorithms automatically group the points into a hierarchy of clusters. Document clustering helps tackle the information overload problem in several ways. One is exploration; the top level of a cluster hierarchy summarizes at a glance the contents of a document collection. Also the features are extracted efficiently.

E. Irrelevant Features and the Subset Selection Problem

We address the problem of ending a subset of features that allows a supervised induction algorithm to induce small high accuracy concepts. We examine notions of relevance and irrelevance_ and show that the definitions used in the machine learning literature do not adequately partition the features into useful categories of relevance. The features selected should depend not only on the features and the target concept_ but also on the induction algorithm. We describe a method for feature subset selection using cross validation that is applicable to any induction algorithm. In this the relevant features alone are extracted.

III.EXISTING SYSTEM

In 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.

A Distortion algorithm, which creates an individual area for each and every word from the already selected transactional database, those are collectively called as dataset, which will be suitable for a set of particular words, but it will be problematic for the set of records.

A Blocking algorithm make propagation to the above problem, and reduce the problems occurred in the existing distortion algorithm, but here also having the problem called data overflow, once the user get confused then they can never get the data back. 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.

Drawbacks of Existing System

- Lacks speed
- Security Issues
- Performance Related Issues
- The generality of the selected features is limited and the computational complexity is large.
- Their computational complexity is low,

IV.PROPOSED SYSTEM

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 but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed FAST algorithm falls into the second group.

Supervised attribute clustering

This algorithm is used to identify uniform cluster that have a high probability density and also class purity will be increase for clustering process. Let C represents the set of attributes of the original data set, while S and S are the set of actual and augmented attribute, respectively, chosen by the proposed attribute clustering algorithm. Let Vi is the coarse cluster related with the attribute Ai and Vi, the finer cluster of Ai, represents the set of attributes of Vi those are merged and averaged with the attribute Ai to generate the augmented cluster representative Ai.

A. Minimum spanning tree

MST is a graph based model in producing the clusters from high computational complexity, it selects or rejects the edges in MST. Spanning tree with their weight less than or equal to the weight of every other spanning tree. Clustering by Minimal Spanning Tree can be view as a hierarchical clustering algorithm which track the divisive clustering approach. Clustering algorithm based on minimum and maximum spanning tree were generally studied to construct MST of point set and delete conflicting edges. Whose weights are expansively larger than the standard weight of the close proximity edges in the tree. The goal to maximize the minimum inters cluster distance.

MST based image segmentation is based on select the edges from the graph, where each pixel correspond to a node in the graph. Weights on every edge calculate the dissimilarity between pixels. The segmentation algorithm define the restrictions between regions by comparing two quantities Intensity difference across the boundary and Intensity difference between neighbouring pixels with all region. This is useful knowing that the intensity differences across the boundary are important if they are huge comparative to the concentration distinction inside the at least on of the regions.

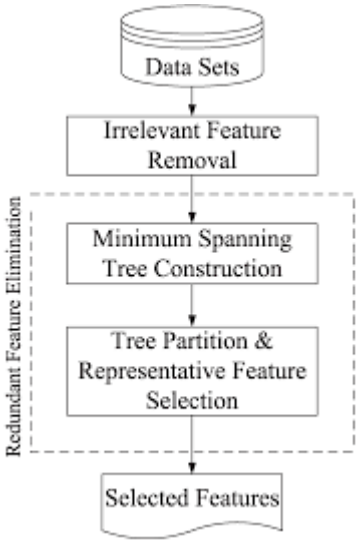

Fig 1. Framework for feature subset selection algorithm

C . Filter method

Filter methods used as a proxy evaluate as an alternative of the error rate to get a feature subset. This measure is elected to be fast to compute. Common procedures in filter methods are Mutual Information, correlation coefficient, and the inter/intra class distance. Filters are usually fewer computationally exhaustive than wrappers, but filter produces a feature set which is doesn't tune to an exact type of predictive model. Many filters afford a feature ranking rather than an explicit best feature subset, and the cutoff point in the ranking is selected via cross-validation.

D.Graph-theoretic clustering

Graph-theoretic clustering are partition vertices in a large graph into different clusters. Both coarse clustering and fine clustering are based on this algorithm called dominant-set clustering. It produces fine clusters on incomplete high dimensional data space. This algorithms that are held to execute well with respect to the indices explain as in the previous section are outlined. The first iteratively emphasise the intra-cluster over inter-cluster connectivity and the second is repeatedly refines an initial partition based on intra-cluster conductance. While together essentially work locally, we also suggest another, more global method. In all three cases, the asymptotic worst-case running time of the algorithms based on certain parameters known as input. However, see that for important choices of these parameters, the time complexity of the novel algorithm GM is superior than for the other two algorithms.

V. Conclusion

An Efficient FAST clustering-based feature subset selection algorithm for high dimensional data improves the efficiency of the time required to find a subset of features. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced and improved the classification accuracy.

References

[1] 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 YEAR 2013.

[2] H. Almuallim and T.G. Dietterich, "Algorithms for IdentifyingRelevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992

[3] M. Dash, H. Liu, Feature selection methods for classification, Intelligent Data Analysis: An Internat. J. 1 (3) (1997).

[4] Forman G., "An extensive empirical study of feature selection metrics for text classification", Journal of Machine Learning Research, 3, pp 1289- 1305,2003

[5] .Huan Liu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE transactions on knowledge and data engineering, VOL. 17, NO. 4, April 2005.

[6] Krier C, Francois D, Rossi F and Verleysen M, "Feature clustering and mutual information for the selection of variables in spectral data", In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, pp 157-162,2007.

[7] .Lei Yu, Huan Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy", Journal of Machine Learning Research, 5, 1205–1224, 2004.

[8] Lei Yu, Huan Liu," Efficiently Handling Feature Redundancy in High Dimensional Data", ACM, August 27, 2003.

[9] Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Learning, 2003.

[10] Lei Yu, Huan Liu," Redundancy Based Feature Selection for Microarray Data", ACM, August 2004.

[11] 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.

[12] Guyon I. and Elisseeff A., An introduction to variable and feature selection,Journal of Machine Learning Research, 3, pp 1157-1182, 2003.

[13] 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.

[14] Jihoon Yang and Vasant Honavar, "Feature Subset Selection Using A Genetic Algorithm Artificial Intelligence Research Group,

[15] H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection: A Filter Solution," Proc. 13th Int'l Conf. Machine Learning, pp. 319-327, 1996.

Author Bibliography

S.Nagendrudu is working as a Lecturer in CSE Department, Satnhiram Engineering College, nandyal, kurnool District,Andhra Pradesh. he has a total of 8 Years Experience in Teaching. he has six International Journal Publications

V.Rama Krishna Reddy is working as a Lecturer in CSE Department, Satnhiram Engineering College, nandyal, kurnool District,Andhra Pradesh. he has a total of 5 Years Experience in Teaching.