Volume 3, Issue 12, December 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Enhanced Mining of High Dimensional Data using Efficient Clustering Algorithm

Suganya.D
PG Scholar
SNS College of Engineering, India

Kowshika.A
Assistant Professor
SNS College of Engineering, India

Abstract- Feature selection in clustering is used for extracting the relevant data from a large collection of data by analyzing on various patterns of similar data. Based on the accuracy and efficiency of the data, major issue occurs in clustering. A novel approach called supervised attribute clustering algorithm is proposed to improve the accuracy and check the probability of the patterns. In this method, faster retrieval of relevant data is made more efficient and accurate. By using this method, users can get precise results and negligible data loss. This method displays results based on the high probability density thereby providing privacy for data and reducing the dimensionality of the data.

Keywords- Supervised Attribute Clustering, Graph Based Method, Filter Method, Finer Cluster.

I. INTRODUCTION

Clustering can be considered as an unsupervised learning since it deals with finding a structure in a collection of unlabeled data, which are similar between them and are dissimilar to the objects belonging to other clusters. This concept is mainly used to simplify the data, detecting the data patterns and identifying features of patterns.The feature of a clustering result depends on both the similarity measure used by the method and its implementation and also measured by its ability to discover some or all of the hidden patterns. Use of Clustering in Data Mining is to identify groups of related records that can be used as a starting point for exploring further relationships. Objects with similar categorical attribute values are placed in the same group and objects in different groups contain dissimilar categorical attribute values.A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. Accuracy and efficiency of data are major issues in clustering. A supervised attribute clustering algorithm is proposed to improve the accuracy and efficiency of data. The main desire of this algorithm is to identify class uniform clusters that have high probability densities and it improves class purity. This method is primarily used for remove redundant data, repeated data and also reducing the dimensionality of data.

II. RELATED WORK

A. Fast Clustering Algorithm

Feature selection can be extracting the feature from the cluster. A Fast clustering bAsed feature Selection algoriThm can divide the features in a cluster using graph theoretic method [4]. This method is partition the vertices in a large graph into different clusters and also detecting densely connected graph into large graphs. To improve the efficiency of fast, MST is used for easy retrieval of features [3]. It also improves performance of classifier.

B. Divisive Information Theoratic Feature

The methods adopted for text classification and word clustering [1]. Distributed clustering applied to a homogenous scenario it cannot ease dimensionality. To overcome this crisis divisive algorithm is used for increasing the value of features and also assesses the effectiveness of the classifier [1]. Model complexity will be reduced and accuracy is high due to using support vector machine.The technique is defined over a vector space where the classification difficulty is to locate the decision surface. It Separate the data points of one class from the other. In crate of linearly separable data, the choice surface is a hyper plane that maximizes the margin between the two classes. Divisive clustering algorithm can be divided into two problems. first, the problem of selecting which cluster must be split then the problem of how splitting the selected cluster.

C. Mutual Information

This technique is used for feature selection of neural network learning using MIFS Algorithm [2].It deals with access the information content from complex classification task. To overcome the problem MIFS algorithm is used to remove the redundant data, irrelevant features and reduce dimensionality [1]. This also maintains the feedback using back propagation algorithms [2]. Greedy technique can evaluate the data content from every individual feature in that each feature is the starting component of a pruning algorithm. This algorithm selects a subset of features from an initial set of available features. Mutual information estimator to grip missing data and to use it to attain feature selection.

D. Feature selection through clustering

The method adopted for feature selection from clustering [1]. This selection process for selecting a subset of features from relevant features, when using the feature selection technique contains many redundant and irrelevant dat.[2] Redundant features , which provide no more information than the selected features, and irrelevant features provide no useful information[3]. Feature selection method is a subset of the more field of feature extraction. Feature extractions create new features from the original set of features, whereas this returns a subset of the features. Feature selection increases the model interpretability, shorter training times by minimizing over fitting. Wrapper methods use a predictive model to achieve feature subsets [3]. Each and every new subset is used to prepare a model, which is tested on a holdout. As wrapper methods train a best model for each subset, these are very computationally rigorous, but it usually provides the better performing feature set for that particular model.

E. Bi-normal Separation

This technique mainly concerns on text classification [1], feature selection is to make huge problems computationally efficient, and storage resources for each and every feature make use of classifiers. To overcome the problem, filter method can be used to remove the irrelevant features and produce a feature set which is not tuned to an exact type of predictive model [3].

F. Analysis of ReliefF and RReliefF

Algorithm adopted for attribute estimators to identify conditional dependencies among attributes and afford attribute estimation in classification. Feature selection is the main problem. To overcome this removes the redundant feature from collecting clusters using relief.

III. PROPOSED METHOD

A. 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 V_i is the coarse cluster related with the attribute A_i and ~~V_i~~, the finer cluster of A_i, represents the set of attributes of V_i those are merged and averaged with the attribute A_i to generate the augmented cluster representative ~~A~~i.

B. 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.

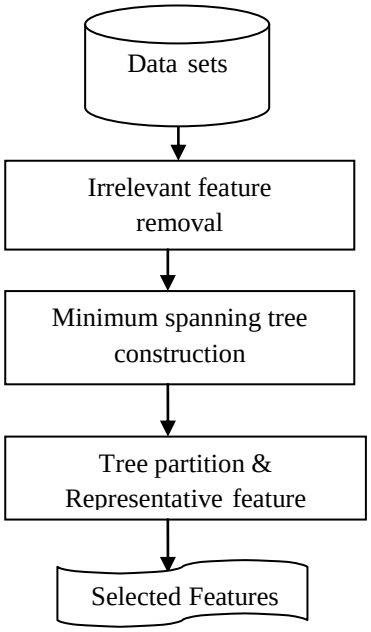

Fig 1. Framework for feature subset selection algorithm

Below diagram shows, Irrelevant features are removed from given data sets, then construct the minimum spanning tree construction and partition the tree to collect the representative features, from that representative features we select the relevant features from MST.

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.

IV. CONCLUSION

The purpose of cluster analysis has been established to be more effective than feature selection algorithms. Since high dimensionality and accuracy are the two major concerns of clustering, we have considered them together in this paper for the finer cluster for removing the irrelevant and redundant features. The proposed supervised clustering algorithm is processed for high dimensional data to improve the accuracy and check the probability of the patterns. Retrieval of relevant data should be faster and more accurate. This displays results based on the high probability density thereby reducing the dimensionality of the data.

REFERENCE

[1] *Qinbao song,Jingjie Ni and Guangtao Wang, "A Fast Clustering Based Feature Subset Selection Algorithm for High-Dimensional Data"IEEE Trans.,vol 25,no.1,January 2013*

[2] *I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification, "J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.*

[3] *R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," July1994.*

[4] *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.*

[5] *G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," J. Machine Learning Research, vol. 3,pp. 1289-1305, 2003.*

[6] *M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of Relief and ReliefF," Machine Learning, vol. 53, pp. 23-69, 2003.*

[7] *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, 2004.*

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

[9] *M. Last, A. Kandel, and O. Maimon, "Information-Theoretic Algorithm for Feature Selection," Pattern Recognition Letters, vol. 22, nos. 6/7, pp. 799-811, 2001.*

[10] *C. Krier, D. Francois, F. Rossi, and M. Verleysen, "Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data," Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning, pp. 157-162,2007*

[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] *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.*

[13] *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.*

