A Review of Feature Selection Techniques for Clustering High Dimensional Structured Data

Bhagyashri A. Kelkar and Dr.S.F. Rodd

Abstract--- Data that resides in a fixed field within a record or file is called structured data and have a defined schema. Structured data is getting more and more importance in database applications such as molecular biology, image retrieval, XML document retrieval etc. The objects are usually represented as a vector of measurements, or a point in multidimensional feature space. To make sense out of the abundance of available information, various data mining and data analysis tools like classification and clustering are being used. However when clustering or classification is done with high dimensional data, traditional algorithms fail to perform as they treat all features equally important in deciding the class/cluster memberships of objects. This is due to the fact that, some of the dimensions are irrelevant and can confuse data mining algorithms by hiding clusters in noisy data. Also in some applications, the cluster structure in the dataset is often limited to a subset of features rather than the entire feature set. Hence feature selection has become an important preprocessing task for effective application of data mining techniques in real-world high dimensional datasets.

Keywords--- Feature Selection, Clustering.

I. INTRODUCTION

DATA mining refers to discovering interesting knowledge from data stored either in databases, data warehouses, or other information repositories. An important technique in data analysis and data mining applications is clustering. From a machine learning perspective clusters correspond to hidden patterns, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics[7,10,11,12], computational biology, and many others. Clustering is a rather diverse topic, and the underlying algorithms depend greatly on the data domain and problem scenario.

Technology advances have made data collection easier and faster, resulting in larger, more complex datasets with many dimensions. The objects are usually represented as a vector of measurements, or a point in multidimensional space. Distance functions lose their usefulness in high dimensionality. In high dimensional data, many of the dimensions are often irrelevant and these irrelevant dimensions can confuse data mining algorithms. Hence it is often valuable to isolate only the most descriptive and discriminatory features in the input set, and utilize those features exclusively in subsequent analysis. Hence feature selection for clustering is an active research topic and important to improve prediction/classification/clustering quality, reduce computation time and build more understandable models. The goal is to improve data mining algorithm performance and computational efficiency. To define interestingness and relevance of features, researchers have proposed measures such as scatter separability, entropy, category utility, maximum likelihood, density, and consensus. Defining interestingness is difficult because it is relative. Given the same data, what is interesting to a physician will differ from what is interesting to an insurance company. Entropy which is a measure of impurity can be used to decide the interestingness of a feature and ensures how much the feature is relevant for classification/clustering purpose.

II. LITERATURE SURVEY

Feature selection can be defined as the process of choosing a minimum subset of M features from the original dataset of N features (M < N), so that the feature space (i.e. the dimensionality) is optimally reduced. An exhaustive search for dominant features would definitely find the optimal solution; however, a search on \(2^N\) possible feature subsets (where \(N\) is the number of features) is computationally impractical. Feature selection methods are categorized as either filter or wrapper approaches [26] based on whether the evaluation methods depend on the learning algorithms. Usually, a wrapper approach may lead to better performance compared to a filter approach for a particular learning algorithm. However, wrapper methods are more computationally expensive since one needs to run the learning algorithm for every candidate feature subset.

A. Filter Methods

Filter methods[21] use some intrinsic property of the data to select features without utilizing the clustering algorithm that will ultimately be applied. The basic components in filter methods are the feature search method and the feature selection criterion. Filter methods are also known as open loop methods because they look only at the intrinsic properties of the data like distance, correlation and consistency. These methods are based on the observation that data with clusters has very different point-to-point distance histogram than that of data without clusters. An entropy measure can be used which is low if data has distinct clusters and high otherwise. The entropy measure is suitable for selection of the most important subset of features because it is invariant with...
number of dimensions, and is affected only by the quality of clustering. Extensive performance evaluation over synthetic, benchmark, and real datasets show its effectiveness.

D. Subspace Clustering

Rakesh Agrawal and his colleagues introduced CLIQUE[25] (Clustering in Quest), a subspace-clustering algorithm that proceeds level-by-level from one feature to the highest dimension or until it generates no more feature subspaces with clusters (regions with high density points). The idea is that dense clusters in dimensionality d should remain dense in d - 1. Subspace clustering also lets to discover different clusters from various subspaces and combine the results. Several new subspace clustering methods were developed after CLIQUE. In clustering, many clusters may exist in different subspaces for small dimensionality with overlapped or non-overlapped dimensions [2]. Subspace searching is not only the feature selection problem. It is finding many subspaces in which feature selection finds one subspace. Therefore, there is a requirement for efficient subspace search algorithms for clustering.

E. Probabilistic Model

Martin H. Law, Anil K. Jain, and Mario A.T. Figueiredo assume that irrelevant features have a probability density identical for all components. So, feature selection and clustering can be performed simultaneously in single expectation-maximization iteration.

F. Co-Clustering

Feature selection can be performed by clustering in the feature space to reduce redundancy. Co-clustering has recently become popular because of research in microarray analysis[12]. Co-clustering is simply clustering the row (sample space) and column (feature space) simultaneously.

G. Other Feature Selection Techniques

Ensemble feature selection [33] is a relatively new technique used to obtain a stable feature subset. A single feature selection algorithm is run on different subsets of data samples obtained from bootstrapping method.

III. METHODOLOGY

There are two major approaches to feature selection. The first is individual evaluation, and the second is subset evaluation. Ranking of the features is known as Individual Evaluation. In Individual Evaluation, the weight of an individual feature is assigned according to its degree of relevance. In Subset Evaluation, candidate feature subsets are constructed using search strategy.

The general procedure for feature selection has four key steps:

1. Subset Generation
2. Evaluation of Subset
3. Stopping Criteria
4. Result Validation

Subset generation is a heuristic search in which each state specifies a candidate subset for evaluation in the search space. First, successor generation decides the search starting point, which influences the search direction. Second, search organization is responsible for the feature selection process with a specific strategy, such as sequential search, exponential search or random search. A newly generated subset must be
evaluated by a certain evaluation criteria. Finally, to stop the selection process, stop criteria must be determined. Feature selection process stops at validation procedure.

While selecting features for clustering, there is a dilemma in choosing which one to start with or which one serves the demands of the other. Should feature selection be utilized to improve clustering quality or clustering is an indicator of relevant features is the question. This leads to the chicken or the egg dilemma i.e. which one comes first, clustering or feature selection. If the answer is feature selection, then feature selection is a goal in itself. However, this is not the case in feature selection literature. Feature selection is used to improve learning quality, reduce computational time and reduce require storage. Thus, the answer to the question is, indeed, feature selection should be used to improve clustering quality. Feature selection that does not depend on any clustering input to define the relevancy of the feature should be preferred. Thus feature selection method should be applied on the whole dataset and then the selected features should be used to construct the clusters. If the clustering quality is not satisfactory, this should be used as an indicator that selected set of features is not satisfactory.

IV. EVALUATION MEASURES FOR FEATURE SELECTION

In order to measure feature selection algorithm’s ability to select relevant features following criteria are used [Dy Brodley] [18].

Cross-Validated Class Error: It is defined as the number of instances misclassified divided by the total number of instances. Each data point is assigned to its most likely cluster, and assign each cluster to a class based on examining the class labels of the training data assigned to each cluster and choosing the majority class. Since true cluster labels are known, classification error can be computed. Class error based on training decreases with an increase in the number of clusters, k, with the trivial result of 0% error when each data point is a cluster. To ameliorate this problem, ten-fold cross-validation error is used. Tenfold cross-validation randomly partitions the data set into ten mutually exclusive subsets.

Recall and precision are concepts from text retrieval (Salton and McGill, 1983) and are defined for feature selection as:

Recall: the number of relevant features in the selected subset divided by the total number of relevant features.

Precision: the number of relevant features in the selected subset divided by the total number of features selected.

These measures give an indication of the quality of the features selected. High values of precision and recall are desired. Finally, to evaluate the clustering algorithm’s ability to find the “correct” number of clusters, average number of clusters found is used as measure.

In [18] Dy. And Brodley have concluded that, standardizing the data before feature subset selection is needed. Finding appropriate no. of clusters (value of k) for given problem leads to better results than fixing k, as different feature subsets have different number of clusters. They also mention that, feature selection obtained better results than without feature selection.

V. CHALLENGES AND FUTURE DIRECTION

Research in feature subset selection for unsupervised learning should be aided with visualization and user interaction to guide the feature search. Another interesting direction is to look at feature selection with hierarchical clustering since hierarchical clustering provides groupings at various perceptual levels.

A. Forward vs Backward Selection

In the literature, it is argued that backward elimination is less efficient than forward selection. Moreover, the computational complexity forward feature selection method is less than backward feature selection. Pros of the forward greedy feature selection method are that it is computationally efficient and does not over fit. Cons, errors made in the early stage by forward greedy feature selection method are do not correct later stages.

B. Feature Selection with Large Dimensional Data

Dimensionality in the range of hundreds is called high dimensional data. Many feature selection algorithms have higher time complexity about dimensionality, therefore the scalability of feature selection is a difficult problem. A filter approach has less computational complexity than a wrapper approach, because it uses independent subset evaluation criteria for subset evaluation. A filter approach is more scalable than the wrapper, so is preferred to a wrapper approach for feature selection. In literature, the embedded approach has been proposed to utilize the qualities of the filter and wrapper high dimension environment. The embedded method has similar time complexity as the filter approach. The inference of the above discussion is that future research must be concentrated on low time complexity with high scalability feature selection algorithms. There is a great research opportunity to develop algorithms using sequential search strategies for clustering.

VI. CONCLUSION

More information is not always good in machine learning applications. For the application at hand, a feature selection algorithm can be selected based on the following considerations: simplicity, stability, number of reduced features, classification accuracy, storage and computational requirements. Overall applying feature selection will always provide benefits such as providing insight into the data, better classifier/clustering model, enhance generalization and identification of irrelevant variables. The efficiency of feature selection techniques is judged by classifier/clustering accuracy and the number of reduced features. A lot of work is being done in supervised feature selection methods mainly using filter evaluation framework. Comparatively little work is done in unsupervised and semi-supervised feature selection methods.
REFERENCES


Bhagyashri A. Kelkar, received the ME in computer sci. & engg. She has 13 years of experience in database based application s/w development and 5 years teaching experience. She is currently working as Head of Dept. of CSE and Assistant professor in Sanjay Ghodawat Group of Institutes Atigre, Kolhapur, Maharashtra, and is a research scholar in Viswaswaraya Technological University, Belagavi, Karnataka. (E-mail-kelkarba@gmail.com)

Dr.S.F. Rodd, is working as Professor and Head of Dept. of CSE in Gogte Institute of Technology, Belagavi, Karnataka(E-mail-sfroodg@it.git.edu)