A Survey on Clustered Feature Selection by using Rough Clustering

Nagalakshmi M
Assistant Professor
Rishi M.S. Institute of Engineering & Technology for Women, Hyderabad ,Andhra Pradesh,India

Dr.Sohan Garg
Director
Sir Chhotu Ram Instituteof Engineering & Technology C.C.S University campus, meerut ,uttar Pradesh,India

ABSTRACT
Feature selection, as a dimensionality reduction technique, aims to choosing a small subset of the relevant features from the original features by removing irrelevant, redundant or noisy features. Feature selection usually can lead to better learning performance, i.e., higher learning accuracy, lower computational cost, and better model interpretability. Recently, researchers from computer vision, text mining and so on have proposed a variety of feature selection algorithms and in terms of theory and experiment, show the effectiveness of their works. In machine learning, feature selection is preprocessing step and can be effectively reduce high dimensional data, remove irrelevant data, increase learning accuracy, and improve result comprehensibility. High dimensionality of data takes over efficiency and effectiveness points of view in feature selection algorithm. Efficiency stands required time to find a subset of features, and the effectiveness belongs to good quality of the subset of features. In feature selection technique high dimensional data contains many irrelevant and redundant features. Irrelevant features make available no useful information in any context, and redundant features provide no more information than the selected features.

1. INTRODUCTION
Recently, available data has increased explosively in both number of samples and dimensionality in many machine learning applications such as text mining, computer vision and biomedical. In order to knowledge acquisition, it is important and necessary to study how to utilize these large scale data. Our interest focus mainly on the high dimensionality of data. The huge number of high dimensional data has imposed significantly big challenge on existing machine leaning methods. Due to presence of noisy, redundant and irrelevant dimensions, they can not only make learning algorithms very slow and even degenerate the performance of learning tasks, but also can lead to difficulty on interpretability of model. Feature selection are capable of choosing a small subset of relevant features from the original ones by removing noisy, irrelevant and redundant features.
In terms of availability of label information, feature selection technique can be roughly classified into three families: supervised methods [1, 2, 3, 4], semi-supervised methods [5, 6], and unsupervised methods. The availability of label information allows supervised feature selection algorithms to effectively select discriminative and relevant features to distinguish samples from different classes. Some supervised methods have been proposed and studied [3]. When a small portion of data is labeled, we can utilize semi-supervised feature selection which can take advantage of both labeled data and unlabeled data. Most of the existing semi-supervised feature selection algorithms [5] rely on the construction of the similarity matrix and select those features that best fit the similarity matrix. Due to the absence of labels that are used for guiding the search for discriminative features, unsupervised feature selection is considered as a much harder problem. In order to attain the goal of feature selection, several criteria have been proposed to evaluate feature relevance.

Wrapper methods use a predetermined learning model to score a feature subsets. A wrapper methods train a fresh model for new subset, they have high accuracy but are expensive to compute and also limited in generality of selected features. Filter methods are faster than wrapper methods but produces a features set which is independent from learning algorithms with better generality. Filter methods measures include the correlation coefficient, Mutual Information, distance and consistency measurements to sort a good subset. Filtering approach to feature selection involves a greater degree of search through the feature space but the accuracy of the algorithms is not guaranteed. Embedded algorithms integrates feature subset selection as a training process and they are fixed to learning methods, hence more efficient than Wrapper and Filter methods. Decision tree algorithms are best example of embedded methods. A combination of filter methods and wrapper methods form hybrid methods which achieves best possible performance with a specific learning algorithm with similar time complexity like the filter methods. The wrapper methods tend to over fit on small training sets. The main benefits of filter methods are they are faster and they have ability to scale to large datasets. With respect to the filter feature selection methods, the application of cluster analysis clearly gives practical

demonstration and explanation to be more effective than traditional feature selection algorithms.

The distributional clustering of words is agglomerative in nature and reduce the high dimensionality of text data since each word cluster can be treated as single feature but are expensive compute. In cluster analysis, most of the applications use a graphtheoretic methods because they produce good results. The graph-theoretic clustering is simple since it compute a neighborhood graph of instances, then delete any edge in graph that is much short or long than its neighbors. The graph theoretic clustering results in forest and trees in forest represents a cluster. In this survey graph-theoretic clustering algorithms are used to features, particularly minimum spanning tree based clustering algorithms.

2. RELATED WORK

2.1. Rough clustering The concept of a rough cluster was introduced by defining a rough cluster in a similar manner to a rough set - with a lower and upper approximation - allowing multiple cluster membership for objects in the data set. The LA of a rough cluster contains objects that only belong to that cluster, and by definition, the objects belong to the UA as well. The UA of a rough cluster contains objects that may belong to more than one cluster. The clustering algorithm described used a distance measure to construct a similarity matrix, and each objectobject pair in this similarity matrix was assigned to existing or new clusters depending on whether none, one or both objects in the pair were currently assigned. Problems with this approach were the large number of clusters generated and uncertainty as to whether the lower approximations of each cluster provide the most efficient coverage of the data set. A different approach was followed in [6], who used reducts to develop clusters. Reducts are subset of the attribute set A, which provide the same information as the original data set. The reducts are used as initial group centroids, which are then grouped together to form clusters. One problem with this approach is that not all information systems have reducts, and some sets of reducts overlap, which means that the cluster centroids are not necessarily well separated. Other approaches include Herawan who used subsets of the total information set to determine a degree of dependency, or the relationship between the subsets, to determine the best clustering attribute, and Yanto who report a variable precision rough set model.

2.2. Comparing rough and k-means clusters Voges reported a comparison of rough clustering with k-means clustering, and found that the two clustering techniques resulted in some clusters that were identified by both techniques, and some clusters that were unique to the particular technique used. The rough clustering solution is necessarily different, because of the possibility of multiple cluster membership of objects. The rough clustering technique also found clusters that were “refined” subclusters of those found by k-means clustering, and which identified a more specific sub-segment of the data set. Rough clustering also produces more clusters than k-means clustering, with the number of clusters required to describe the data dependent on the distance measure. More clusters means an object has a higher chance of being in more than one cluster. A solution with too few clusters does not provide a useful interpretation of the partitioning of the data. On the other hand, too many clusters make interpretation difficult. In addition, the degree of overlap between the clusters needed to be minimized to ensure that each cluster provided information to aid in interpretation. Rough clustering can be conceptualized as extracting concepts from the data, rather than strictly delineated sub-groupings. Determining a good rough cluster solution requires a trade-off between various factors. As we show below, evolutionary algorithms are a good way of conducting this trade-off.

2.3. Evolutionary algorithms and rough sets

A number of applications of evolutionary algorithms to rough clustering tasks have been reported in the literature. Mitra proposed an evolutionary rough cmeans clustering algorithm to determine the relative importance of upper and lower approximations of rough sets used to model the clusters. The fitness function used in the evolutionary algorithm involved minimizing a specific measure, the Davies–Bouldin clustering validity index. Kumar used an agglomerative hierarchical clustering algorithm for sequential data, where the indiscernibility relation was extended to a tolerance relation with the transitivity property being relaxed. Bouyer [3] used a Kohonen self-organizing map for pre-processing of data, which was then further divided into clusters using rough sets and genetic algorithms. How the genetic algorithm was applied is not clearly described, but it appears to use a data structure based on inter-neuron distances in the selforganizing map. As this distance measure is based on Euclidean distances, the approach is restricted to continuous attributes. Lingras developed a genome comprising two sections – LA membership and UA membership. The approach required some repair operators, as some randomly generated genes could be invalid. One limitation of this approach was that the number of clusters needed to be specified in advance, and this preliminary knowledge is not always available for larger data sets. There have also been a number of applications of evolutionary algorithms to classification tasks using rough sets. For example, [5]

used a hybrid system to develop linguistic-based technical stock market indicators with rough sets theory used to extract linguistic rules and a genetic algorithm to refine these extracted rules. The effectiveness of the proposed model was verified for both forecasting accuracy and stock returns, and showed that the proposed model was superior to rough sets and genetic algorithms applied independently. Salamó proposed several rough set based measures for estimating attribute relevance for feature dimensionality reduction in Case-Based Reasoning classifiers.

3. CLUSTER ANALYSIS

Cluster analysis is a fundamental technique in both traditional data analysis and in data mining. The technique is defined as grouping 'individuals or objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters' [8:470]. Many clustering methods have been identified, including partitioning, hierarchical, nonhierarchical, overlapping, and mixture models. One of the most commonly used nonhierarchical methods is the k-means approach. In the k-means approach, objects are randomly selected as initial seeds or centroids, and the remaining objects are assigned to the closest centroid on the basis of the distance between them. The aim is to obtain maximal homogeneity within subgroups or clusters, and maximal heterogeneity between clusters. The data set is partitioned into clusters and an error term e is calculated, usually based on the Euclidean distance between each object and the cluster centroids. The usual approach is to search for a partition with small e by moving cases from one partition to another. The search through the problem space to find the lowest value of e is considered computationally expensive and local optimization has traditionally been used. In addition, the number of clusters in each partition is decided prior to the analysis, a major limitation of the technique.

A k- Modes approach has been developed as an extension of the k-Means algorithm, and has been applied to categorical data clustering by replacing means with modes [4]. However even with this extension, the number of clusters needs to be set in advance. In the last few decades, as data sets have grown in size and complexity, and the field of data mining has matured, many new techniques based on developments in computational intelligence have started to be more widely used as clustering algorithms. For example, the theory of fuzzy sets developed by Zadeh introduced the concept of partial set membership as a way of handling imprecision in mathematical modeling. This was subsequently applied to cluster analysis [2],

One technique from the field of computational intelligence receiving considerable attention is the theory of rough sets. In most previous applications of rough sets theory, the technique was used for classification problems, where prior group membership is known, and results are usually expressed in terms of rules for group membership. This paper describes a rough clustering technique, based on a simple extension of rough sets theory, applicable where prior group membership is not known. Before describing this technique, a brief introduction to canonical rough sets theory is provided.

4. TYPES OF CLUSTERING ALGORITHMS FOR HIGHDIMENSIONAL DATA SPACE

In this section, we describe some of the clustering algorithms for High Dimensional data space. These are specific and need more attention because of high dimensionality [2]. Today, most of the research work is carrying under this. Due to high dimensionality it is becoming tedious and needs more generalized techniques to cluster various dimensions of the data [3]. Due its dimensionality, there is a need for dimensionality reduction and redundancy reduction at the time of clustering. This section discusses the main subspace clustering and projected clustering strategies and summarizes the major subspace clustering and projected algorithms [6].

SUBSPACE CLUSTERING

Subspace clustering methods will search for clusters in a particular projection of the data. These methods can ignore irrelevant attributes and also problem is known as Correlation clustering. Two-way clustering, or Co-Clustering or Biclustering are known as the special case of axis-parallel subspaces. In these methods the objects are clustered simultaneously as the feature matrix consisting of data objects as they are span in rows. As in general subspace methods they usually do not work with arbitrary feature combinations. But this special case it deserves attention due to its applications in bioinformatics.

CLIQUE-

Clustering in Quest, is the fundamental algorithm used for numerical attributes for subspace clustering. It starts with a unit elementary rectangular cell in a subspace. If the densities exceeds the given threshold value, those cell are will be retained [5]. It applies a bottom-up approach for finding such units. First, it divides units into 1- dimensional equal units with equal-width bin intervals as grid. Threshold and bin intervals are the inputs for this algorithm. It uses Apriori-Reasoning method as the step recursively

from q-1-dimensional units to q-dimensional units using selfjoin of q-1. The total subspaces are sorted based on their coverage. The subspaces which are less covered are pruned. Based on MDL principle a cut point is selected and a cluster is defined as a set of connected dense units. A DNF expression that is associated with a finite set of maximal segments called regions is represented whose union is equal to a cluster [6].

PROJECTED CLUSTERING

Projected clustering tries to assign each point to a unique cluster, but the clusters may exist in different subspaces. The general approach uses a special distance function along with a regular clustering algorithm. PROCLUS -Projected Clustering, [2], is associates with a subset of a lowdimensional subspace S such that the projection of S into the subspace is a tight cluster. The pair (subset, Subspace) will represent a projected cluster. The number of clusters k and average subspace dimension n will be specified by the user as inputs [6]. It finds k-medoid in iterative manner and each medoid is associated with its subspace. A sample of data is used along with greedy hill-climbing approach and the Manhattan distance divides the subspace dimension. An additional data passes follow after the iterative stage is finished to refine clusters with subspaces associated with the medoids. ORCLUS-Oriented projected Cluster generation [3] is an extended algorithm of earlier proposed PROCLUS. It uses projected clustering on non-axes parallel subspaces of high dimensional space.

HYBRID CLUSTERING ALGORITHM

Sometimes it is observed that not all algorithms try to find a unique cluster for each point nor all clusters in all subspaces may have a result in between. It is because of having a number of possibly overlapping points. The exhaustive sets of clusters are found necessarily. FIRES [4], can be used as a basic approach a subspace clustering algorithm. It uses a heuristic aggressive method to produce all subspace clusters.

CORRELATION CLUSTERING

Correlation Clustering is associated with feature vector of correlations among attributes in a high dimensional space. These are assumed to persistent to guide the clustering process [2]. These correlations may found in different clusters with different values, and cannot be reduces to traditional uncorrelated clustering [6]. Correlations among attributes or subset of attributes results different spatial shapes of clusters. Hence, the local patterns are used to define their similarity between cluster objects. The Correlation clustering can be considered as Biclustering as both are related very closely. In the biclustering, it will identify the groups of objects correlation in some of their attributes. The correlation is typical for the individual clusters.

5. EXPERIMENT RESULTS

We give the clustering results of different methods on the 12 real life datasets in Table 1(ACC) and Table 3(NMI).

The results include the average and the standard deviation of clustering accuracy and normalized mutual information, respectively. From the two tables, we can make the following several observations. First, feature selection is necessary and effective. It can not only significantly reduce the numbers of feature and make machine learning algorithms more efficient, but also can improve the performance. Secondly, in general, almost no one feature selection method can obtain the best result on all data sets.

Table 1. Clustering results of different methods on 12 data sets. The best result for each data set is highlighted in bold face.

Dataset	ACC ± std(%)								
	All Features	MaxVar	Laplacian Score	SPFS-SFS	SPEC	MCFS	UDFS	NDFS	EUFS
PIE10P	26.7 ± 1.5	27.1 ±1.1	30.1 ± 0.4	28.9 ± 2.1	27.5 ± 0.8	29.3 ± 2.1	29.5 ± 3.3	29.4 ± 1.6	**47.5 ± 2.3**
PIX10P	85.2 ± 3.3	82.9 ± 3.6	86.9 ± 4.7	86.2 ± 3.2	86.1 ± 5.2	**88.1 ± 6.7**	83.6 ± 2.9	83.3 ± 7.5	86.5 ± 4.0
COIL20	62.7 ± 3.1	61.4 ± 1.6	62.2 ± 1.9	64.3 ± 2.1	65.5 ± 3.8	65.9 ± 2.2	65.5 ± 2.9	63.9 ± 2.4	**66.2 ± 2.7**
ORL	49.7 ± 3.2	50.8 ± 1.4	49.9 ± 2.4	50.4 ± 1.2	51.4 ± 2.2	57.0 ± 3.2	53.8± 3.0	**57.6 ± 1.7**	50.2 ± 2.3
JAFFE	85.3 ± 6.1	85.5 ± 4.2	86.2 ± 3.7	87.1 ± 3.3	85.9 ± 5.1	90.7 ± 6.1	90.5 ± 1.4	**91.0 ± 3.4**	80.1 ± 6.2
MNIST	51.8 ± 2.0	52.0 ± 1.7	52.6 ± 1.8	54.1 ± 1.1	52.4 ± 0.5	52.2 ± 0.3	**57.1 ± 1.2**	49.6 ± 1.1	53.2 ± 2.1
BA	40.9 ± 1.6	41.7 ± 1.3	43.3 ± 1.9	43.9 ± 1.4	42.7 ± 1.1	42.9 ± 1.8	43.8 ± 1.6	42.9 ± 1.8	**45.6 ± 1.4**
tr11	31.8 ± 2.2	31.4 ± 2.4	**39.5 ± 3.2**	37.6 ± 1.2	38.0 ± 3.1	32.1 ± 1.8	35.5 ± 2.1	34.6 ± 1.4	35.5 ± 1.9
oh15	31.6 ± 2.7	32.2 ± 2.1	34.7 ± 2.4	**35.2 ± 1.9**	34.2 ± 2.0	32.5 ± 1.3	32.6 ± 2.4	34.5 ± 1.7	34.2 ± 1.9
TOX-171	42.8 ± 2.1	42.9 ± 1.6	43.1 ± 1.4	44.5 ± 0.3	40.4 ± 0.0	42.9 ± 1.6	45.6 ± 1.2	**46.9 ± 1.5**	42.0 ± 1.8
Tumors9	40.8 ± 3.7	41.2 ± 2.6	42.3 ± 2.6	42.9 ± 2.7	35.8 ± 2.4	42.4 ± 3.6	43.3 ± 3.5	**45.6 ± 4.6**	42.2 ± 3.9
Leukemia1	61.0 ± 5.9	61.3 ± 4.2	62.5 ± 0.0	79.2 ± 2.1	81.6 ± 1.6	69.7 ± 3.2	81.0 ± 3.8	**90.5 ± 2.5**	72.5 ± 4.2

Table 2. Clustering results of different methods on 12 data sets. The best result for each data set is highlighted in bold face.

Dataset	NMI ± std(%)								
	All Features	MaxVar	Laplacian Score	SPFS-SFS	SPEC	MCFS	UDFS	NDFS	EUFS
PIE10P	25.5 ± 3.4	28.6 ± 2.7	30.5 ± 2.5	30.8 ± 0.5	25.3 ± 1.5	31.9 ± 3.1	**49.9 ± 2.7**	30.1 ± 3.1	49.3 ± 1.8
PIX10P	88.0 ± 2.1	89.1 ± 1.6	89.8 ± 0.7	90.0 ± 3.2	91.0 ± 1.9	**91.7 ± 3.1**	85.6 ± 1.9	86.8 ± 4.5	91.5 ± 1.3
COIL20	77.1 ± 1.3	71.9 ± 0.7	72.5 ± 1.1	73.7 ± 0.5	75.3 ± 1.6	74.5 ± 1.2	76.0 ± 1.3	74.3 ± 1.8	**76.6 ± 1.7**
ORL	70.0 ± 1.7	70.7 ± 2.1	71.1 ± 1.3	70.9 ± 1.2	71.4 ± 1.3	75.2 ± 1.7	73.4±1.5	**75.6 ± 1.6**	70.5 ± 1.3
JAFFE	87.5 ± 3.8	83.1 ± 3.4	87.2 ± 2.4	90.8 ± 3.7	87.4 ± 2.2	**91.4 ± 3.8**	90.3 ± 5.2	89.4 ± 2.1	82.3 ± 3.4
MNIST	48.9 ± 1.0	47.6 ± 0.4	48.1 ± 1.0	48.9 ± 0.4	48.3 ± 0.4	**52.0 ± 0.2**	50.0 ± 0.9	44.8 ± 0.5	47.5 ± 0.7
BA	57.2 ± 1.1	57.7 ± 0.9	58.7 ± 0.7	58.9 ± 1.2	58.3 ± 0.8	58.6 ± 0.8	**59.1 ± 0.9**	58.1 ± 0.9	58.4 ± 0.9
tr11	5.7 ± 1.6	8.9 ± 2.2	15.2 ± 3.5	**15.3 ± 3.4**	14.5 ± 3.0	7.1 ± 1.7	11.1 ± 1.6	9.9 ± 3.5	12.7 ± 3.9
oh15	20.5 ± 2.1	23.2 ± 1.6	25.7 ± 1.9	**26.2 ± 1.3**	24.9 ± 1.6	23.4 ± 1.1	23.2 ± 2.1	22.3 ± 1.8	24.5 ± 2.7
TOX-171	13.6 ± 2.3	11.4 ± 3.2	12.5 ± 1.7	20.2 ± 3.2	9.7 ± 0.0	12.7 ± 0.4	16.7 ± 4.8	**22.3 ± 1.8**	13.0 ± 1.7
Tumors9	39.5 ± 3.1	40.2 ± 2.5	41.0 ± 2.3	41.3 ± 2.1	34.5 ± 2.4	41.1 ± 2.7	41.5 ± 3.5	**44.1 ± 3.4**	41.1 ± 3.2
Leukemia1	37.6 ± 10.7	36.1 ± 5.5	36.7 ± 0.0	49.3 ± 2.6	58.5 ± 1.7	53.5 ± 1.3	59.6 ± 4.5	**66.2 ± 7.4**	61.8 ± 0.9

CONCLUSIONS

This paper gives a survey on feature selection methods proposed in literature. Several state of the art feature selection methods are introduced. As we can see in our experiments, there are one or more parameters to be set. The purpose of this article is to present a comprehensive classification of different clustering techniques for high dimensional data. Clustering high dimensional data sets is a ubiquitous task. The incosent growth in the fields of communication and technology, there is tremendous growth in high dimensional data spaces. It study focuses on issues and major drawbacks of existing algorithms. As the number of dimensions increase, many clustering techniques begin to suffer from the curse of dimensionality, de-grading the quality of the results. However, in practice, we do not and can not know the best parameters corresponding to the given data set. So How to select the adaptive hyper-parameters and the number of selected features are open problems and also are our future work.

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] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279-305, 1994.

[3] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.

[4] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.

[5] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103, 1998.

[6] Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537-550, 1994.