International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 11 (2017) pp. 2893-2899
© Research India Publications. http://www.ripublication.com

A Survey on Clustering High Dimensional Data Techniques

M. Pavithra[1], and Dr. R.M.S.Parvathi [2.]

[1]*Assistant Professor, Department of Computer Science and Engineering, Jansons Institute of Technology, Coimbatore, India.*

[2]*Dean- PG Studies, Sri Ramakrishna Institute of Technology, Coimbatore, India.*

[1]ORCID: 0000-0003-0412-6711

Abstract

Dimension reduction is commonly defined as the process of mapping high-dimensional data to a lower-dimensional embedding. Applications of dimension reduction include, but are not limited to, filtering, compression, regression, classification, feature analysis, and visualization. We review methods that compute a point-based visual representation of high-dimensional data sets to aid in exploratory data analysis. The aim is not to be exhaustive but to provide an overview of basic approaches, as well as to review select state-of-the-art methods. The main objective of this research paper is to prove the effectiveness of high dimensional data analysis and different algorithm in the prediction process of Data mining. The approach made for this survey includes, an extensive literature search on published papers as well as text books in the application of Data mining in prediction. I have retrieved those articles by searching the data bases with the usage of the keywords "data mining and algorithm". Titles of the articles were analysed by usage of association rules that analyse the most frequently used words. The main algorithm which were includes in this survey are decision tree, k-means algorithm, and association rules. Therefore an analysis of the different classes of available clustering techniques with big datasets may provide significant and useful conclusions. The proposed system is to study and analyse some of the popular existing clustering techniques and impact of dimensionality reduction on Big Data. The objective of the proposed model is to improve the performance of traditional H-K clustering and overcome the limitations such as high computational complexity and poor accuracy for high dimensional data by combining the three different approaches of clustering algorithm as subspace clustering algorithm and ensemble clustering algorithm with H-K clustering algorithm. Clustering high dimensional datasets is a monotonous task due to the curse of dimensionality. The main drawback of k means clustering is that the accuracy of the clusters fully depends on the selection of initial centroids. Initialization methods are suggested by researchers to improve the performance of k-Means clustering performances. But these methods do not provide adequate results for clustering high dimensional data. In this paper, a novel approach for clustering high dimensional data collected from the Facebook is proposed. Clustered Facebook data is used to find the closeness between two participants in the social network. Initially, the dimensions are reduced using modified Principal Component Analysis (PCA) called "Reduced Uncorrelated Attributes (RUA)". Then the transformed dataset is used to find the initial seed using maximum-Average-Minimum (k-MAM) method for k-means. The performance of the proposed method is analysed and the result shows that RUA with k-MAM provides more accuracy than other methods.

Keywords: Data Mining, Clustering, High Dimensional data, Clustering Algorithm, Dimensionality Reduction.

INTRODUCTION

Clustering is a technique in data mining which deals with huge amount of data. Clustering is intended to help a user in discovering and understanding the natural structure in a data set and abstract the meaning of large dataset. It is the task of partitioning objects of a data set into distinct groups such that two objects from one cluster are similar to each other, whereas two objects from distinct clusters are dissimilar [1]. Clustering is unsupervised learning in which we are not provided with classes, where we can place the data objects. Clustering is beneficial over classification because cost for labelling is reduced. Clustering has applications in molecular biology, astronomy, geography, customer relation management, text mining, web mining, etc. All applications use clustering to derive useful patterns from the data which helps them in decision making This is helpful to draw certain conclusions and proceed further in that direction for enhancement of application.

Cluster Analysis is an important tool for exploratory data analysis which aims at summarizing main characteristics of data. Lot of work has been done in past in this field. Clustering algorithms such as K-mean algorithm has a history of fifty years [2]. Clustering methods have also been developed for categorical data. Clustering methods are applied in pattern recognition [3], image processing and information retrieval. Clustering has a rich history in other disciplines [4] such as biology, psychiatry, psychology, archaeology, geology, geography, and marketing [5]. Cluster finding methodologies differ from need to need of an application. Clustering methods and algorithms are dependent on number of instances to be considered, size of a single instance, accuracy of the result required. All these factors give rise to various methods and algorithms. This paper presents an overview of clustering techniques, their comparison, advantages and disadvantages of them, and challenges that needs attention. Section II describes clustering methods, section III of paper discusses how clustering differs when it is high dimensional data. Section IV focuses on subspace clustering. In section V we compare various methods and finally discuss applications and conclusion in the last two remaining sections.

Clustering is a division of data in to groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification [9]. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modelling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis [5]. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data [10].

Clustering can be considered as the most important unsupervised learning problem. Clustering deals with finding a primitive structure in a collection of unlabelled data. A cluster a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [2]. The objective of the clustering technique is to determine the intrinsic grouping in a set of unlabelled data. The similarity between data objects can be measured with the imposed distance values. Specifying the distance measures for the high dimensional data is becoming very trivial because it holds different data values in their corresponding attributes[4]. Following is the analysis of different distance measures used for measuring similarity between data objects in clustering [6].

Distance Measure: Most of the clustering techniques relay on distance measure as an important step for selecting data objects, which will determine the similarity between two elements [2]. The cluster shape or density will be influenced by the similarity between the data objects [3], as some elements may be close to one another according to one distance and farther away according to another. In general there are two types of distance measures. 1) Symmetric measure and 2) Asymmetric measure. The common distance measures used in the clustering process [3][4] are i)The Euclidean distance or Squared Euclidean distance, ii)The Manhattan distance , iii)The Maximum Likelihood Distance, iv)The Mahalanobis distance,(v)The Hamming distances ,(vi) The angle between two vectors used as a distance measure when clustering high dimensional data.

LITERATURE SURVEY

The field of Data Mining has a different behaviour towards Big Data. It can deal with data-sets having size gigabytes or even tera bytes. The main concern over here is that the algorithms which are used in data mining operations work on small data sets and do not give better results on large data sets. To work efficiently with large data sets, the algorithms must have high scalability. Clustering high dimensional data has always been a challenge for clustering techniques. Clustering is unsupervised classification of patterns (observations, data items, or feature vectors) into teams (clusters). The drawbacks of clustering have been addressed in several contexts by researchers in several disciplines and so reflect its broad charm and quality in concert of the steps in exploratory data analysis. Clustering is useful in several exploratory pattern analysis, grouping, decision making and machine learning situations including data mining, document retrieval, image segmentation and pattern classification. Adil Fahad, etal[1] performed a survey on clustering algorithms for Big Data. They have categorized 24 Clustering Algorithms as Partition-based, Hierarchical-based, Density based, Grid-based and Model-based. Depending on the size of datasets, handling capacity of noisy data and types of datasets, Clusters are formed and the complexities of algorithms are calculated. They concluded that no clustering algorithm performs well for all the evaluation criteria. The entire clustering algorithm suffers from Stability problem. MacQueen [7] defined a technique for partitioning N-dimensional population into k-sets, which they named as K-means. They successfully concluded that k-means is computationally feasible and economical and has been a successful implementation for differentiating the data within a class. S.Nazim presented a comparative review of dimensionality reduction techniques in regard with information visualization. The survey analysed some DR methods supporting the concept of dimensionality reduction for getting the visualization of information with minimum loss of original information. As, we are dealing with Big Data. The issue of stability of clusters comes into picture. The theories [8] state that k-means does not break down even for arbitrarily large samples of data. The focus is on the behaviour of stability of clusters formed by k-means algorithm-Means is closely related to principal component analysis [9]. The outcomes subject with regard to effectiveness of the solution obtained from k-means. Unsupervised dimensionality reduction and unsupervised learning are associated closely [10]. The result provides new perception towards the observed quality of output obtained by PCA-based data reduction.

Bara'a Ali Attea et al. [6] discovered that performance of clustering algorithms degrades with more and more overlaps among clusters in a data set. These facts have motivated to develop a fuzzy multi-objective particle swarm optimization framework (FMOPSO) in an innovative fashion for data clustering, which is able to deliver more effective results than state-of-the-art clustering algorithms. To ascertain the superiority of the proposed algorithm, number of statistical tests has been carried out on a variety of numerical and categorical real life data sets.

Suresh Chandra Satapathy et al. [7] introduced an idea of an algorithm that can combine dimensionality reduction techniques of weighted PCs with AUTO-PSO for clustering. The intention behind it was to reduce complexity of data sets and speed up the Auto PSO clustering process. A significant improvement in total runtime has been achieved. Moreover, the clustering accuracy of the dimensionality reduction technique i.e. AUTO-PSO clustering algorithm is comparable to the one that uses full dimension space.

Li-Yeh Chuang et al. [8] invented an improved particle swarm optimization based on Gauss chaotic map for clustering. Gauss chaotic map provides the significant chaos distribution to balance the exploration and exploitation capability for search process. This easy and fast function generates a random seed processes, and further improve the performance of PSO due to their unpredictability. In the experimental analysis, the eight different clustering algorithms were compared on six test data sets. The results indicated that the performance of the proposed one is appreciably better than the performance of other existing algorithms.

Xiaohui Cui et al. [9] presented a Particle Swarm Optimization (PSO) document clustering algorithm. Unlike, localized searching of the K-Means algorithm, PSO clustering algorithm used to perform a globalized search in the entire solution space. In the experiments conducted, they have applied the K-Means PSO, and hybrid PSO clustering algorithm on four different text document data sets. From the comparative results, the hybrid PSO algorithm can generate more compact clustering results than the K-Means algorithm.

CLUSTERING HIGH DIMENSIONAL DATA

The objects in data mining could have hundreds of attributes. Clustering in such high dimensional

Spaces present tremendous difficulty, much more so than in predictive learning. In decision trees, for example, irrelevant attributes simply will not be picked for node splitting, and it is known that they do not affect Nave Bayes as well. In clustering, however, high dimensionality presents a dual problem. First, under whatever definition of similarity, the presence of irrelevant attributes eliminates any hope on clustering tendency [11]. After all, searching for clusters where there are no clusters is a hopeless enterprise. While this could also happen with low dimensional data, the likelihood of presence and number of irrelevant attributes grows with dimension. The second problem is the dimensionality curse that is a loose way of speaking about a lack of data separation in high dimensional space. Mathematically, nearest neighbour query becomes unstable: the distance to the nearest neighbour becomes indistinguishable from the distance to the majority of points [10]. This effect starts to be severe for dimensions greater than 15.

Therefore, construction of clusters founded on the concept of proximity is doubtful in such situations. For interesting insights into complications of high dimensional data see [Aggarwal et al. 2000]. Basic exploratory data analysis (attribute selection) preceding the clustering step is the best way to address the first problem of irrelevant attributes [8]. We consider this topic in the section General Algorithmic Issues. Below we present some techniques dealing with a situation when the number of already pre-selected attributes is still high .In the sub-section Dimensionality Reduction we talk briefly about traditional methods of dimensionality reduction [4]. In the sub-section Subspace Clustering we review algorithms that try to circumvent high dimensionality by building clusters in appropriate subspaces of original attribute space. Such approach has a perfect sense in applications, since it is only better if we can describe data by fewer attributes. Still another approach that divides attributes into similar groups and comes up with good new derived attributes representing each group is discussed in the sub-section Co-Clustering [9]. Important source of high dimensional categorical data comes from transactional (market basket) analysis. Idea to group items very similar to co-clustering has already been discussed in the section Co-Occurrence of Categorical Data [1].

ANALYSIS OF HIGH DIMENSIONAL DATA FOR CLUSTERING

The rapid growth in various new application domains, like bioinformatics and e-commerce, reflects the need for analysing high dimensional data. Many organizations have massive amounts of data containing valuable information for running and building a decision making system [2]. To do this, it makes study and to analyse high dimensional and large amount data for effective decision making. Generally, in a gene expression microarray data set, there could be tens or hundreds of dimensions, each of which corresponds to an experimental condition. Researchers and practitioners are very eager in analysing these data sets [3]. However, before analysing the data mining models, the researcher will analyse the challenges of attribute selection, the curse of dimensionality, redundancy reduction, data labelling and the specification of similarity in high dimensional space for analysing high dimensional data set [5].

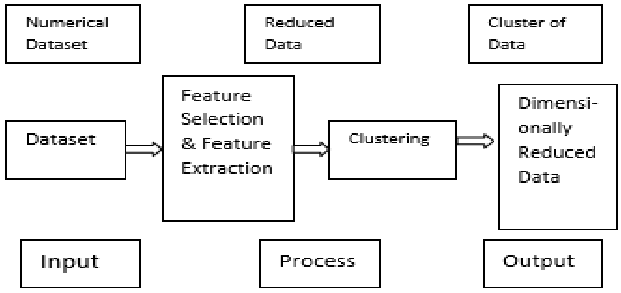
In data mining, the objects can have hundreds of attributes or dimensions. Clustering in such high dimensional data spaces presents a tremendous difficulty, much more so than in predictive learning [6]. In clustering, however, high dimensionality presents two problems.

1) The clustering tendency will lose when the dataset contains irrelevant attributes [5]. Searching for clusters is a hopeless enterprise where there are no relevant attributes for finding clusters. Attribute selection is the best approach to address the problem of selecting irrelevant attributes.

2) Dimensionality curse is another problem in high dimensional data. As the number of attributes or dimensions increases in a dataset, the distance measures will become increasingly meaningless [6] [7]. The resultant clusters with high dimensions; they are equidistant from each other.

DIMENSIONALITY REDUCTION

Dimensionality curse is a loose way of speaking about lack of data separation in high dimensional space [7], [6], and [8]. The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions [7]. With increasing dimensionality, these algorithms soon become computationally intractable and therefore inapplicable in many real applications. In general, there are two approaches that are used for dimensionality reduction. One is attribute Transformation and another one is attribute Decomposition. Attribute Transformations are simple function of existent attributes. For example, the sales profiles and OLAP-type data, rollup as sums or averages over time intervals can be used. In multivariate attribute selection can be carried out by using Principle Component Analysis (PCA) [9]. Attribute Decomposition is a process of dividing data into subsets. Using some similarity measures, so that the high dimensional computation over smaller data sets will happen [10]. Dimensions stay the same, but the costs are reduced. This approach targets the situation of high dimensions, large data.



It was proven that, for any point in a high dimensional space, the expected gap between the Euclidean distance to the closest neighbour and that to the farthest point shrink as the dimensionality grows [8]. This phenomenon may render many data mining tasks ineffective and fragile because the model becomes vulnerable to the presence of noise. An Adaptive dimension reduction for clustering, a new semi-supervised clustering framework based on feature projection and fuzzy clustering is proposed for clustering high dimensional data [11]. In this proposed model, the standard practice of reporting the results directly obtained in the reduced-dimension subspace is not accurate enough.

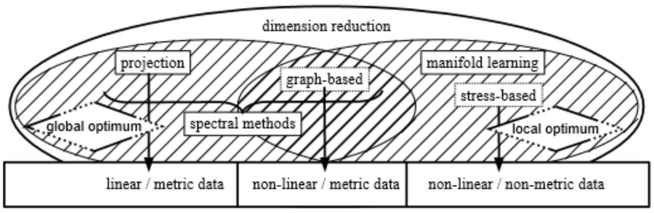
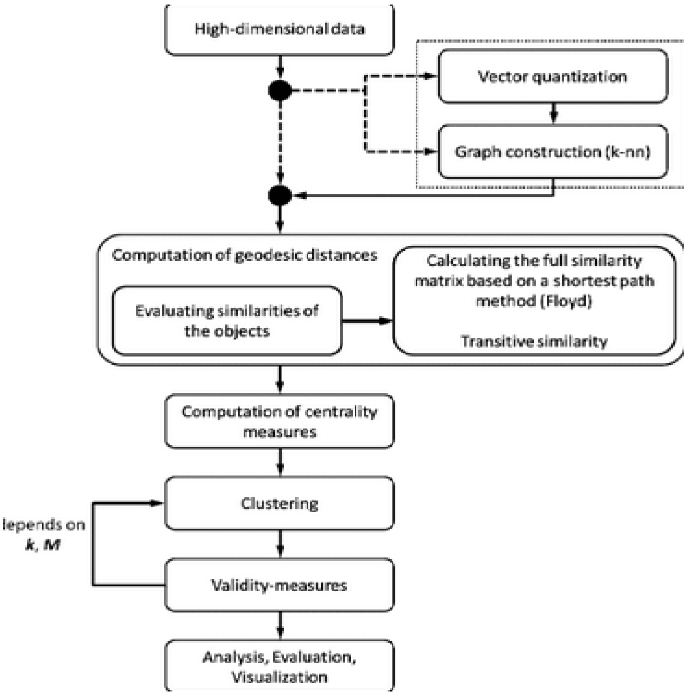


Figure 1: Concepts of dimension reductiuon

CLUSTERING ALGORITHMS FOR HIGH DIMENSIONAL DATA

The main aspiration of clustering is to find high quality clusters within reasonable amount of time. Clustering in data mining is the process of discovering groups [2]. Each group is a dataset such that the similarity among the data inside the group is maximized and the similarity in outside group is minimized. The discovered clusters are then used to explain the characteristics of the data distribution [4]. Today there is tremendous necessity in clustering the high dimensional data. For example, many business applications, clustering can be used to describe different customer groups and allows offering customized solutions [6]. Clustering can be used to predict customer buying patterns based on their profiles to which cluster they belong. In the following section 4 presents various types of clustering algorithms used for two dimensional data space and section 5 represents types clustering algorithms high dimensional data space[7].


High-dimensional data
Vector quantization
Graph construction (k-nn)
Computation of geodesic distances
Evaluating similarities of the objects
Calculating the full similarity matrix based on a shortest path method (Floyd)
Transitive similarity
Computation of centrality measures
Clustering
lepends on k, M
Validity-measures
Analysis, Evaluation, Visualization

TYPES OF CLUSTERING ALGORITHMS FOR HIGH-DIMENSIONAL 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 [12]. 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 and [11]. 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 [13], 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 [8]. It uses Apriori-Reasoning method as the step recursively from q-1-dimensional units to q-dimensional units using self-join 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 low-dimensional 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 [9].

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 [7]. 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 [9].

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 [8]. 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 [10].

CLUSTERING HIGH DIMENSIONS DATA TECHNIQUES

The operation of clustering high dimensional data techniques has recently grown in advance. The popular methods as mentioned above were analysed in detail.

GAUSSIAN MIXTURE MODELS USING HIGH-DIMENSIONAL DATA

Clustering divides a given dataset {x1... xn} of n data points into k homogeneous groups. Popular clustering techniques use Gaussian Mixture Models (GMM), which assume that each class is represented by a Gaussian probability density. Data k{x1... xn} ∈ Rp are then modelled with the density f (x, θ) = ∑where φ is a multi-variate normal density with parameter θi = {μi , Σi } and πi are the mixing proportions[7]. This model which uses to estimates full covariance matrices and therefore the number of parameters is very large in high dimensions. However, due to the empty space phenomenon we can assume that high-dimensional data live in subspaces with a dimensionality lower than the dimensionality of the original space [4]. We here propose to the work in low dimensional class specific subspaces in order to adapt classification to high dimensional data and to limit the number of parameters to estimate [6].

THE DECISION RULE

Classification assigns an observation x ∈ Rp with unknown class membership to one of k classes C1... Ck knew a priori. The optimal decision rule is the one which called Bayes decision rule, this affects the observation x to the class which has the maximum posterior probability P (x ∈ Ci |x) = πi φ(x, θi)/ l=1 πl φ(x, θl). Maximizing the posterior probability is equivalent to minimizing −2 log (πi φ(x, θi)). For the model [aij bi Qi di], this results in the decision rule δ + which assigns x to the class minimizing the following cost function Ki (x)

$$K_i(x)=\|\mu_i-P_i(x)\|^2_{A_i}+\frac{1}{b_i}\|x-P_i(x)\|^2+\sum_{j=1}^{d_i}\log(a_{ij})+(p-d_i)\log(b_i)-2\log(\pi_i)$$

Where ||.||Λi is the Mahalanobis distance associated with the matrix Λi =Qi ΔI Qit. The posterior probability can therefore be rewritten as follows: to identify dubiously classified points. We can observe that this new decision rule is mainly based on two distances: the distance between the projection of x on Ei and the mean of the class; and the distance between the observation and the subspace Ei . This rule assigns a new observation to the class for which it is close to the subspace and for which its projection on the class subspace is close to the mean of the class [1]. If we consider the model [ai bi Qi di], the variances ai and bi balance the importance for the both distances. The example, if the data having too much noisy, i.e. bi is large, it is natural to balance the distance ||x − Pi (x)||2 by 1/bi in order to take into account the large variance in E(1/i) .Remark that the decision rule δ+ of our models uses only the projection on Ei and we only have to estimate a di - dimensional subspace [4]. Thus, our models are significantly more parsimonious than the general GMM. For example, if we consider 100-dimensional data, that are made of 4 classes with common intrinsic dimensions di equal to 10, the model [ai bi Qi di] requires the estimation of 4 015 parameters whereas the full Gaussian mixture model estimates 20 303 parameters [7].

HIGH DIMENSIONAL DATA CLUSTERING

In this section we derive the EM-based clustering framework for the model [aij bi Qi di] and the sub-models. The new clustering approach are referred to by the High-Dimensional Data Clustering, which has the lack of space, we do not need to present the proofs of the following results which can be found in [9].

THE CLUSTERING METHOD HDDC

Unsupervised classification organizes data in homogeneous groups using only the observed values of the p, whereas p is the explanatory variables. Normally, the parameter uses to estimate by the EM algorithm which repeats iteratively E or M steps [10]. Suppose if we use the parameterization that presented in the previous section, that the EM algorithm for estimating the parameters θ = {πi , μi , Σi , aij , bi , Qi , di }, would be written as follows:

E STEP: this step computes at the iteration q the conditional posterior probabilities: tij(b)=P(xj£Ci(q)|xj),from the relation , it may consider:

tij(q)=1/∑

(1/2(Ki(q-1)(xj)-Kl(q-1)(xi)))

where Ki is defined

M STEP: this step maximizes at the iteration q has the conditional likelihood. Proportions, which means and covariance matrices of the mixture are estimated by: πi(q)=(ni(q) /n), μi(q)=(1/ni(q))∑

ij (q)xj,ni(q)=∑

ij (q) (2) ∑ (1/ni(q))∑

ij (q)(xj-μi(q))(xj-μj(q))t

The estimation of the HDDC parameters is detailed in the following subsection.

ESTIMATION OF HDDC PARAMETERS

Assuming for the moment that parameters di is known and omitting the index q of the iteration for the sake of simplicity, we obtain the following closed form estimators for the parameters of our models:–

SUBSPACE Ei : the di is the first columns of Qi, that are estimated by the eigenvectors associated with the di largest eigenvalues λij of Σi .

Model [aij bi Qi di]: the estimators of aij that having the di largest eigenvalues λij of Σi and the estimator of bi is the mean of the (p − di) smallest eigenvalues of Σi and can be written as follows: bi= (Tr(∑i)-∑

ij) (4)

Model[ai bi Qi di]: the estimator of bi which given at (4) and the estimator of ai is the mean of the di largest eigenvalues of ∑i Ai=(1/di)∑ ij

We also have to estimate the intrinsic dimensions of each subclass. That is very difficult problem which has no unique technique to use. Our approach is based on the eigenvalues of the class conditional where the covariance matrix Σi of the class is Ci [10]. Whereas the jth eigenvalue of Σi corresponds to the fraction of the full variance carried by the jth eigenvector of Σi .Therefore we estimated the class specific dimension di , i = 1, 2,3,4.., k, with the empirical method screen-test of Cattell [3] which analyses the differences between eigenvalues in order to find a break in the screen. The selected dimension is the one for where the subsequent differences are smaller than the threshold. In our experiments, the threshold is chosen by the cross-validation. We also compared the probabilistic criterion BIC which gave very similar results [8].

CONCLUSION

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 [11]. 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. In high dimensions, data becomes very sparse and distance measures become increasingly meaningless. This problem has been studied extensively and there are various solutions, each appropriate for different types of high dimensional data and data mining procedures [12]. There are many potential applications like bioinformatics, text mining with high dimensional data where subspace clustering, projected clustering approaches could help to uncover patterns missed by current clustering approaches. As with any clustering techniques, finding meaningful and useful results depends on the selection of the appropriate clustering technique [5]. In order to do this, one must understand the dataset in a domain specific context in order to be able to best evaluate the results from various approaches. From the above discussion it is observed that the current techniques will suffers with many problems [1]. To improve the performance of the data clustering in high dimensional data, it is necessary to perform research in the areas like dimensionality reduction, redundancy reduction in clusters and data labelling [3].

The principal challenge for clustering high dimensional data is to overcome the “curse of dimensionality”. There are several recent approaches to clustering high dimensional data. These approaches have been successfully applied in many areas [8]. We need to compare these different techniques and better understand their strengths and limitations. A particular method can be suitable for a particular distribution of data. We cannot expect that one type of clustering approach will be suitable for all types of data or even for all high dimensional data. Many issues like scalability to large data sets, independence of order of input, validating clustering result are resolved to much extent [7]. We need to focus on methods which can give us result in a manner which is easy to interpret. Result obtained should be in a manner which can also give us some conclusion and information about data distribution. It should further suggest us on how the clusters obtained can be helpful for various applications [4].

REFERENCES

[1] P. Berkhin, “A Survey of Clustering Data Mining Techniques” Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press, 25-72, 2011.

[2] Guha S., Rastogi R., Shim K,”CURE: An efficient clustering algorithm for large databases”, Proc. Of ACM SIGMOD Conference, 2012.

[3] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann Publishers, 2010.

[4] A. K. Jain and R. C. Dubes, “Algorithms for Clustering Data”, Prentice Hall, 2009.

[5] A. Jain, M. N. Murty and P. J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Volume 31(3), pp. 264-323, 2011.

[6] Zhang T., Ramakrishnan R. and Livny M,” BIRCH: An efficient data clustering method for very large databases”, In Proc. of SIGMOD96, 2012.

[7] Rui Xu and W. Donald, "Survey of Clustering Algorithms," IEEE Transaction on Neural Network, vol. 16, 2009.

[8] Gan Guojan, Ma Chaoqun, and W. Jianhong,” Data Clustering: Theory, Algorithm and Applications”, Philadelphia, 2012.

[9] A. Jain and R. Dubes, “Algorithms for Clustering Data”, New Jersey, 2011.

[10] A. K. Jain, M. N. Murtyand, and P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys vol. 31, pp. 264-324, 2012.

[11] K. Bache and M. Lichman. (2013). UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml/machinelearning-databases/

[12] M. Steinbach, L. Ertoz, and V. Kumar, "The Challenges of Clustering High Dimensional Data," in New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition, Ed New Vistas: Springer, 2010.

[13] Z. Tian, R. Raghu, and L. Miron, "BIRCH: A New Data Clustering Algorithm and Its Applications," Data Mining and Knowledge Discovery, vol. 1, pp. 141-182, 2009.