ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

Clustering Algorithms for High Dimensional Data – A Survey

Maithri.C[1], Chandramouli.H[2]

Assoc. Professor, Research Scholar, Department of Computer Engineering, Kalpataru Institute of Technology, Tiptur, Karnataka, India[1]

Professor, Head, Department of Computer Engineering, EPCET College of Engineering, Bangalore, Karnataka, India[2]

ABSTRACT: 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. With the advent growth of high dimensional data such as microarray gene expression data, and grouping high dimensional data into clusters will encounter the similarity between the objects in the full dimensional space is often invalid because it contains different types of data. The process of grouping into high dimensional data into clusters is not accurate and perhaps not up to the level of expectation when the dimension of the dataset is high. 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 performance issues of the data clustering in high dimensional data , it is also necessary to study issues like dimensionality reduction, redundancy elimination, subspace clustering, co-clustering and data labelling for clusters are to analysed and improved. In this paper, we presented a brief comparison of the existing algorithms that were mainly focusing at clustering on high dimensional data**.**

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

I. INTRODUCTION

Clustering is a technique in data mining which deals with huge amount of data. it is intended to help a user in discovering and understanding the natural structure in a data set and abstract the meaning of large dataset. Clustering is an unsupervised learning in which we are not provided with classes, where we can place the data objects and it is beneficial over classification because cost for labeling is reduced. Clustering has applications in molecular biology, astronomy, geography, customer relation management, text mining, web mining, etc.

Cluster Analysis is an important tool for exploratory data analysis which aims at summarizing main characteristics of data. Clustering techniques can be used to discover natural groups in data sets and to identify a structure that might reside there, without having any specific background knowledge as characteristics of the data. Clustering has been used in a variety of areas, including computer vision, VLSI design, psychology, data mining, bioinformatics, statistics, pattern recognition, machine learning and information retrieval. The objective of the clustering technique is to determine the intrinsic grouping in a set of unlabeled 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. Following is the analysis of different distance measures used for measuring similarity between data objects in clustering.

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. 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. 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.

II. LITERATURE SURVEY

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.

Bara'a Ali Attea et al. 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. 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.

Xiaohui Cui et al. 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.

III. ANALYSIS OF HIGH DIMENSIONAL DATA FOR CLUSTERING

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 analyzing these data sets. 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. In clustering, however, high dimensionality presents two problems.

1. 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.

A. *Dimensionality Reduction:*

The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions. With increasing dimensionality, these algorithms soon become computationally intractable and therefore inapplicable in many real applications.

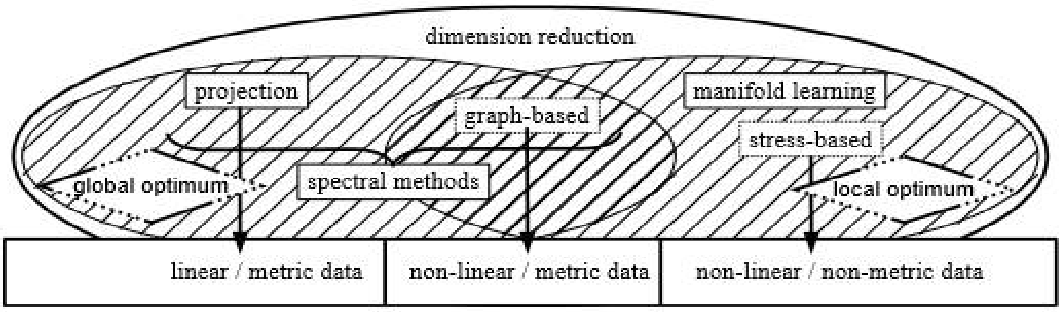

Fig.1: Concepts of dimension reduction.

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. In this proposed model, the standard practice of reporting the results directly obtained in the reduced-dimension subspace is not accurate enough

B. *Sub Space Clustering*

A data object may be a member of multiple clusters, each one existing in a different subspace. In general the Subspace clustering techniques involve two kinds of approaches. One is projection pursuit clustering assumes that the class centers are located on same unknown subspaces. On the other hand, principal component clustering assumes that each class is located on an unknown specific subspace. If the subspaces are axis-parallel, a finite number of subspaces are possible. Hence, subspace clustering algorithm utilizes a kind of heuristics to remain computationally feasible, at the risk of producing inferior results.

C. *Co-Clustering*

Grouping attribute in conjunction with clustering of data points themselves is called co-clustering. Co-clustering will improve clustering of points based on their attributes. It tries to cluster attributes based on their points. Grouping rows and columns into point-by-attribute data representation is the concept of co-clustering. Co-clustering is also known as simultaneous clustering, conjugate clustering, distributional clustering, bi-dimensional clustering, block clustering, and information bottleneck method. Co-clustering on categorical data is prominent area for research now a days. In the Co-Occurrence of Categorical Data the similar way of building groups of items was presented.

IV. Types Of Clustering Algorithms For High- Dimensional Data Space

Most of the research work is carried under this domain. Due to high dimensionality it is becoming tedious and need more generalized techniques to cluster various dimensions of the data. 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.

A. *Subspace Clustering:*

These methods can ignore irrelevant attributes and also problem is known as Correlation clustering. Two-way clustering, or Co-Clustering or Bi -clustering 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. 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 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.

B. *Projected Clustering:*

PROCLUS -Projected Clustering is associated 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. 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 is an extended algorithm of earlier proposed PROCLUS. It uses projected clustering on non-axes parallel subspaces of high dimensional space.

C. *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, can be used as a basic approach a subspace clustering algorithm. It uses a heuristic aggressive method to produce all subspace clusters.

D. *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. These correlations may found in different clusters with different values, and cannot be reduces to traditional uncorrelated clustering. 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.

V. TYPES OF CLUSTERING ALGORITHMS FOR TWO- DIMENSIONAL DATA SPACE

The clustering algorithms for two dimensional data space are specified below. The general categories of the clustering algorithms are listed below

A. K-means clustering algorithm:

The traditional clustering algorithm is the k-means algorithm . In k-means it assigns each point to the cluster which is nearer to the center called centroid. The center is the average of all the points in the cluster that means the coordinates are the simple arithmetic mean for each dimension separately over all the points in the cluster. Simplicity and speed is the main advantage of this algorithm. It also allows running on large datasets. The disadvantage is that at each run it does not produces the same result, since the resulting clusters depend on randomly initialized assignments. The problem by seeking to choose better starting clusters is addressed by k-means. The intra-cluster variance is minimized, but it does not sure about minimizing global variance. Another disadvantage is the requirement of mean to be definable for the concept which the case is not always. For such datasets the variant of k-mean is k-medoid is appropriate. A different criterion for which points are best assigned to which centre are k-medians is clustering.

B. Hierarchical Clustering Algorithm:

Hierarchical clustering builds cluster hierarchy or it's a tree of clusters. It finds successive clusters using previously established clusters. These algorithms can be agglomerative or divisive . Agglomerative hierarchical clustering is a bottom-up clustering. It begins with each element as a separate cluster and merges them into successively larger clusters. Divisive hierarchical clustering is top-down clustering. It's clustering starts with everybody in one cluster and ends up with everyone in individual clusters. Divisive algorithms begin with a set and keep on dividing it into successively smaller clusters.

C. Density-Based Clustering Algorithm:

The data space is divided into a set of its connected components. The basic idea for partitioning into sets requires a concept of density. A cluster defined as a dense component, where it can grow in any direction that density leads. These are devised to discover arbitrary-shaped clusters. In this approach, a cluster is considered as a density region in which the data objects exceeds a threshold. Since it requires space as metric for clustering Density based algorithms are also called as Spatial Data Clustering.

D. Quality Threshold Clustering Algorithm:

Quality Threshold clustering algorithm is an alternative method of partitioning data, particularly invented for gene clustering . It requires excessive computing power than k-means, and does not require the number of clusters in advance. But it always returns the same result as it runs for several times. In QT-Clustering algorithm , the distance between a point and a group of points is computed using complete linkage, which is the maximum distance from the data point in a group to any member of the group.

VI. RESULT SECTION

The section starts by explaining the details of the used methods in our survey of high dimensional data.

A. Dimensionality Reduction:

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.

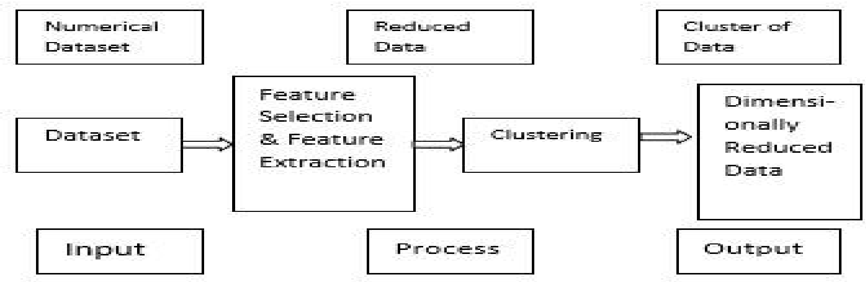

Fig. 2: Dimension Data Set Process

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 .In this proposed model, the standard practice of reporting the results directly obtained in the reduced-dimension subspace is not accurate enough

B. Cluster Method:

In many business applications, clustering can be used to describe different customer groups and allows offering customized solutions. Clustering can be used to predict customer buying patterns based on their profiles to which cluster they belongs.

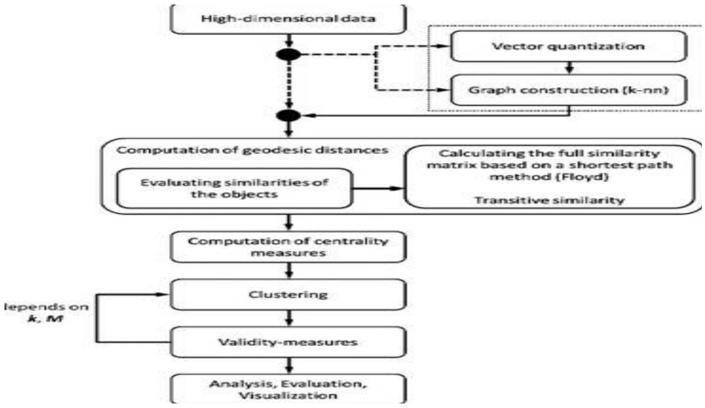
High-dimensional data
Vector quantization
Graph construction (k-nn)
Computation of geodesic distances
Evaluating similarities of the objects
Calculating the full similarity matrix based on a shortest path method (Floyd)
Transitive similarity
Computation of centrality measures
lepends on k, M
Clustering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

Fig.3: Clustering for High-Dimension Data

VII. CONCLUSION AND FUTURE WORK

The innocent growth in the fields of communication and technology, there is tremendous growth in high dimensional data spaces. Its 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. 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.

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, independent of order of input, validating clustering result are resolved to much extent. 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.

REFERENCES

1. Paul E Green and Vithala R Rao, " A Note on Proximity Measures and Cluster Analysis", in Journal Of Marketing Research, 359-64, 1969.
2. Mark A. Hall, Geo_rey Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", IEEE Transactions on Knowledge and data engineering, VOL. 15, NO. 3, MAY/JUNE 2003.
3. Beyer, k., Goldstein, j., Ramakrishnan, r., and Shaft, U. " When is nearest neighbor meaningful?", In Proceedings of the 7th ICDT, Jerusalem, Israel., 1999.
4. Aggarwal, C.C., Hinneburg, A., and Keim, D.A. "On the surprising behavior of distance metrics in high dimensional space", . IBM Research report, RC 21739, 2000.
5. McCullum. A., Nigam, K., and Ungar, L.H." Efficient clustering of high dimensional data sets with application to reference matching". In proceedings of the 6th ACM SIGKDD, 167-178, Boston., MA, 2000.
6. Chris Ding Xiaofeng He, "Adaptive dimension reduction for clustering high dimensional data ", in the Proceedings of IEEE International Conference on Data Mining, Washington DC, USA, 2002.
7. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Volume 31(3), pp. 264-323, 2011.
8. Gan Guojan, Ma Chaoqun, and W. Jianhong," Data Clustering: Theory, Algorithm and Applications", Philadelphia, 2012.