ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.5

Analysis of High Dimensional Data Clustering

Laxmi Sharma[1], Arko Bagchi[2]
M.Tech Student[1], Associate Professor[2]
Department of Computer Science
Delhi College of Technology and Management, India

Abstract:
The mapping process of high dimensional to a lower dimensional data is known as "dimension reduction". Feature analysis, regression, compression, visualization, filtering are included in dimension reduction applications but these applications are not limited to these features only. The dissemination of distances between data points in low dimension looks different than in high dimension set of data. The phenomena of concentration of norms is suffered by the data sets in high dimension. Under some usual hypothesis the distance between the data points in high dimensions is expected to be similar whether the data point is closest or farthest to its neighborhood. In high dimension the analysis of clusters is mainly done on the basis of distances. This paper discussed about the effects of some properties over high dimensional data.

Keywords: Dimension reduction, real valued attributes, high dimensional data, dendogram, unsupervised learning, curse of dimensionality, concentration of norms.

I. INTRODUCTION

Clustering is a technique for analyzing big data applied to the sets of data so that groups of data can be discovered. Data entities with in the same group share similar features and in different group data objects are different. It shows that concept of analysis of clustering is based on similarity of feature or closeness. Here high dimensional data clustering is focused having real valued attributes. The term curse of dimensionality is referred when the circumstances occurred quite often in which large number of dimensions are needed to be handled. In many algorithms in cluster analysis the dimensions number does not affect computational performance. Instead of number of dimensions the number of data points are quite more critical for computational performance. Curse of dimensionality covers all general features of high dimensional data and not just criticality in computational complexity. Euclidean distance is used to measure distance for high dimensional data gives different properties if used for low dimensional data. There are two examples from these types of properties which are – the method of concentration of norms and hubness method. The concentration of norms states that for higher dimensional data distance from any point to its neighbor whether nearest neighbor or farthest neighbor is almost similar. In hubness method, the dispersion of occurrence of data points in between data points k nearest neighbor which are skewed to the right. Hubs are founds very rarely in between other data points k nearest neighbor. In hubness method a count is set for the occurrence of each entity in data points in between k nearest neighbor of other data points. Here k is a fixed constant. The hubness method is a boundary effect and not direct effect of high dimensional data. In cluster analysis, the data objects are grouped into clusters which are similar in same cluster and different in other cluster group. The data having real-valued attributes the similarity in data entities is measured via the distance measures like Manhattan or Euclidean distance. Each data entity is treated as a separate clusters and then in each step the nearest point is merged for the formation of large clusters till no data entity is left. Distance measures are divided into two types which are symmetric and asymmetric measures. Dendogram is used to decide the number of clusters. Dendogram defines the order in which clusters are merged and the distance between the merged clusters. Hierarchical clustering lies in the category of relational clustering techniques which is based on the distance matrix. In this distance matrix the paired distances of the data entities are kept. The number of data points cause distance matrix quadratic in nature and this quadratic complexity with in the data points cannot be avoided in hierarchical clustering. As clustering is a type of unsupervised learning, classes to place data objects are not provided. Due to reduction of labelling cost clustering is advantageous over classification. Clustering is used in customer relation management, molecular biology, geography, astronomy, web mining, text mining etc. All clustering application gives patterns which are derived from data objects which is helpful in decision making. Cluster analysis is a tool for data analysis whose objective is to summarize the main features of data. Clustering techniques and algorithms have dependency over number of instances, accurate results and size of a single distance. Due to all these factors several clustering techniques and algorithms are introduced. According to machine learning clusters are hidden patterns, cluster search is an unsupervised learning which represents data concept. Therefore clustering of hidden data lies in category of unsupervised learning.

II. LITERATURE SURVEY

Clustering gives different results in case of big data sets. In this case it has to deal with dataset having gigabytes or even terabytes of size. Here the main disadvantage is that clustering algorithm gives efficient results in small data sets but their performance reduces in case of large data set which is main concern here. Algorithm should be scalable if the efficient performance of the algorithm is required. High dimensional data clustering is proved to be a challenge for all clustering techniques. Clustering is applied in grouping, data mining in machine learning, decision making, pattern analysis, image segmentation, document retrieval and pattern classification. The survey of clustering algorithm for big data was performed by Adil Fahad, etal. They divide clustering algorithms in 24 categories as Hierarchical based, Grid based, density based,

model based and partitioned based. Clusters formation depends on the dataset size, data set types and capacity of handling noisy data and thus calculation of complexity in performance of algorithms are calculated. The problem of stability is suffered by all clustering algorithm.

There are many methods which apply global dimensionality reduction and use standardizes clustering techniques. Feature selection or feature extraction techniques are the basis for dimension reduction. New variables were developed through feature extraction which contains most of the part of global information. A linear technique known as principal component analysis method are best known for feature extraction [1].

The combination of global feature selection and model based clustering techniques was proposed in a recent approach [2]. The technique to divide N-dimensional population into K-sets is defined by MacQueen and these k-sets are termed as K-means [3]. They made the conclusion that K-means is quite feasible in computation and economical as well. When many overlap occurs in clusters in between a data set then the efficiency of clustering techniques decreases. This was introduced by Barra Ali Attea [4].

III. HIGH DIMENSIONAL DATA CLUSTERING

In clustering, the data objects have hundreds of attributes for high dimensional data. In predictive learning difficulties is caused by presence of spaces. Irrelevant attributes in decision trees are not picked for splitting of nodes which do not affect Nave Bayes. High dimensionality causes two problems in clustering which are:

- Irrelevant attributes in clustering are present which causes removal of hope of clustering tendency and also clsuters are searched when there are no clusters present which is quite hopeless. Although this can also happen in low dimensional data but the presence of irrelevant attributes increases with the increase in dimensions.

- The next problem is curse of dimensionality which is due to very less separation of data in high dimensions. Due to which the distance of closest neighbor can not be separated to the distance of majority of points. For dimensions greater than 15, this effect becomes more severe.

There are some approaches which partition the attributes into several groups and introduce new good attributes. Transaction analysis derive important source for categorical high dimensional data.

Dimension Reduction:

The curse of dimensionality says that data in high dimensional space there is very less separation of separation in data [5]. As dimensionality increases the computational complexity increase and therefore application of algorithms in this case becomes intractable. Two techniques used in dimensionality reduction which are attribute decomposition and attribute transformation [6]. The function of existing attributes are known as simple function. The selection in multivariate attribute is done by applying Principle Component Analysis. In the process of attribute decomposition the data is partitioned into subsets by applying measures of similarity so that computation of smaller data sets in high dimension will takes place.

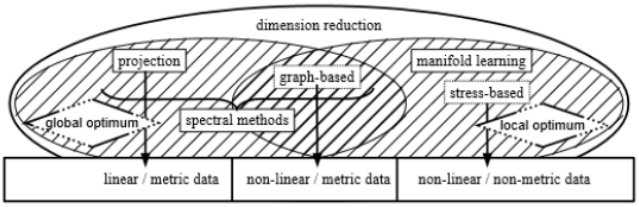

Figure.1. Dimension reduction concepts

The gap between Euclidean distance to the nearest point and to the extreme point decreases as the dimensionality increases. This process can cause clustering techniques to be feeble because the developed model become weak due to the presence of noise. The main objective of clustering is to form high quality of clusters with in less amount of time. The formed clusters explain features of distribution of data.

Clustering Algorithm for High Dimensional Data:

In this section some clustering algorithms which can be applied in high dimensional space will be discussed. There are certain types of clustering which are:

Subspace Clustering: This is an extension of traditional clustering and it looks for data in a specific projection of data i.e. in different subspace with in a dataset [7]. Subspace clustering methods can avoid irrelevant attributes and the problem caused is called as Correlational Clustering. CLIQUE-Clustering is an example of algorithm used for subspace clustering and that to for numerical attributes only.

Projected Clustering: Projected clustering allocate each point to a particular but unique cluster. These clusters can present in distinct subspace. It may be possible that on all the dimensions clusters are not necessarily defined due to deficiency of data but always there will be some subsets of dimensions which can always present on some high quality clusters. These subsets of dimensions can change over distinct clusters. These type of clusters are called as projected clusters [8].

Hybrid Clustering: Hybrid clustering is a clustering method that divides the set of data into primary clusters and then build a hierarchical structure upon these sub clusters which is based on various measures of similarity. This algorithm is based on K-Means and K-Harmonic Means. The K-Means algorithm is of type partitional clustering algorithm. KHM is a center based algorithm which uses averages of distances from each data point to the centers as a component to its performance function. All subspace clusters are produced by using heuristic aggressive method [9].

Correlation Clustering: In high dimensional space the feature vector of correlations among attributes is associated with correlation clustering. These correlations are searched in distinct clusters and cannot be decreased to traditional uncorrelated clustering. Attributes of subsets or correlations between the attributes gives distinct spatial shapes of clusters.

IV. CONCLUSION

In clustering high dimensional data the challenge is to overcome the curse of dimensionality. There are various current techniques to cluster high dimensional data. It is highly needed compare these techniques and to comprehend their limitations and strengths. A specific method suited to a specific data distribution. It cannot be expected that a single type of clustering techniques will be appropriate for all types of data. Many issues like independency in input order, scalability in wide datasets and clustering validation can be resolved at much extent. Focus on those technique is highly needed which results in easy interpretation. The obtained result should be able to provide information about distribution of data as well

as conclusion. It should also suggest the formation of clusters in various applications.

V. REFERENCES

[1] Charles Bouveyron, Stephane Girard, and Cordelia Schmid – "High Dimensional Data Clustering" https://pdfs.semanticscholar.org/e61c/9758a320a8733b75618c25f11234ac9a5395.pdf

[2] C. Bouveyron, S. Girard and C. Schmid – "High-Dimensional Data Clustering" https://arxiv.org/ pdf/math/0604064.pdf

[3] M. Pavithra1, and Dr. R.M.S.Parvathi – "A Survey on Clustering High Dimensional Data Techniques" https://www.ripublication.com/ijaer17/ijaerv12n11_38.pdf

[4] Tian Zhang, Raghu Ramakrishnan and Miron Livny – "Birch – An Efficient Data Clustering Method for Large Database"https://www.cs.sfu.ca/CourseCentral/459/ han/ papers / zhang96.pdf

[5] Huan Xu, Shie Mannor and Constantine Caramanis - "Robust dimensionality reduction for high-dimension data" https://ieeexplore.ieee.org/document/4797709/

[6] Seung-Hee Bae, Jong Youl Choi, Judy Qiu and Geoffrey C. Fox – "Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation" http://citeseerx .ist.psu.edu/ viewdoc/download?doi=10.1.1.154.9672&rep=rep1&type=pdf

[7] Lance Parsons, EhteSham Haque and Huan Lieu – "Sup space Clustering For High Dimensional Data: A Review "http:// www.kdd.org/exploration_files/parsons.pdf

[8] Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu – "A Framework for Projected Clustering of High Dimensional Data Streams "http://web.engr.illinois.edu/~ hanj/pdf/vldb04_projclstream.pdf

[9] Jacob Kogan, Charles Nicholas, Mike Wiacek – "Hybrid clustering of large dimensional data"http://www.math.umbc.edu/~kogan/technical_papers/2006/Kogan_Nicholas_Wiacek.pdf