A SURVEY ON CLUSTERING ALGORITHMS FOR HIGH DIMENSIONAL DATA

Jijo Varghese[1] *Dr. P.TamilSelvan*[2]

ABSTRACT

Clustering in data mining works with a large volume of data. Clustering leads the customer to uncover and understand the standard synthesis of the information set and exemplifies the motive behind an enormous dataset. This research paper aims to focus on the widely used clustering algorithms for sorting and classifying big data. It is essential for analysts to understand the way data are classified for presenting insights into business decisions. Performance issues of data clustering, while simultaneously taking care of high-dimensional data, are discussed including the learning of issues, reduction in dimensionality and disposal, subspace clustering, co-grouping and information marking for groups. Here, a concise study of the present algorithms is talked about, for the most part, focusing on the high dimensional data grouping.

Keywords: Data mining, Dimensionality Reduction, Clustering, High dimensional data and Big data analytics.

1 INTRODUCTION

Data Mining sorts all the big data sets to collect valuable information. During the mining process, computational software recognizes patterns and builds relationships to resolve problems. Companies adopt it for predicting future trends. More popular data mining techniques are required due to the size of the information in Big Data, along with a wide range of data and superfast speed for delivery. Of the different data mining techniques that include association, classification, clustering, organization, prediction, outlier analysis, we have chosen to focus on clustering. Huge amount of data are managed by Clustering. The technique of Clustering helps the users discover and identify the natural structure of the data set. In the unsupervised learning named as Clustering, there is no provision of classes where the data can be placed, and several benefits are provided over classification due to the reduction in cost for labeling. Clustering has been utilized in various areas such as sub-atomic science, space science, topography, client connection, content mining, web mining, and so on. Cluster Analysis plays a significant role in providing exploratory data analysis aiming to summarize the prime characteristics of data. Clustering procedures will find characteristic gatherings in the data sets. Clustering has found its applications in a range of regions, ranging from PC vision to VLSI structure, neuroscience, bioinformatics, machine learning, insights, data mining, pattern recognition and data recovery. The imposed distance values measure similarities between data objects. Defining the distance measures for the high dimensional data is necessary as it holds various kinds of data values in their corresponding attributes. Each meeting is known as a cluster, consisting of relative items. Communicating data to fewer organizations loses specific fine subtleties regardless of the degree of disarray. Research papers on clustering of data and data models are very rare. The performance of data

[1] *Research Scholar, Department of CS, CA&IT Karpagam Academy of Higher Education, Coimbatore*

[2] *Research Supervisor, Department of CS, CA&IT, Karpagam Academy of Higher Education, Coimbatore*

characterizes the clustering of recorded concepts discovered in science, measurement and numerical inquiry. According to machine learning, organizations contrast with shrouded models, where the search for clusters is unaided learning, and the following structure speaks of the concept of data.

2 LITERATURE REVIEW

Lopez (2018) analyzed the problem of cluster analysis for high dimensional data. Some algorithms including the basic and advanced specific solutions were analyzed to overcome the problem. K-means has shown the best results while competing with DBSCAN and COPAC for the tests conducted to carry out the work.

The cluster analysis was elaborated by Steinbach et al. (2004) highlighting the issues associated with clustering data of high dimensionality. A classification was presented by Babu et al. (2011) for the numerous clustering methods of data having high dimensionality. The curse of dimensionality troubled several clustering techniques with the rising number of dimensions and hence the quality of the results declined. A clustering-based feature selection algorithm by Elankavi et al. (2017) was a very fast Clustering Algorithm to select the subset of features. The proposed algorithm incorporated certain features, including the elimination of unnecessary features, development of a minimum spanning tree along with dividing the MST and selection of the representative characteristics. A clustering method by Bouveyron et al. (2007) proposed to estimate the specific subspace and intrinsic dimension for each class. The Gaussian mixture model was chosen for handling data of high dimensionality, and the best-fit parameters for the data were estimated. For locating the objects, the High Dimensional Data Clustering (HDDC) was developed in the original images through the help of a probabilistic framework.

Pavithra &Parvathi (2017) presented as comprehensive categorization of various kinds of clustering techniques for high dimensional data.

An overview was shown by Parsons et al. (2004) for different subspace grouping algorithms incorporating a progressive system arranging the algorithms through their features. The experimental adaptability, along with precision tests, was used to perform a comparison between the two primary techniques with subspace clustering, where several known applications are discussed.

A diagram was presented by Consent (2012) to show the impact of spaces with high dimensionality for several types of ideal clustering models. An evaluation was carried out on models working upon high measurement through clustering, markers to study the literature and design the research challenges prevalent in the current scenario.

3 Types of clustering algorithms for high dimensional space

Most of the research work is performed in this domain. Clustering different dimensions of data is becoming tedious due to high dimensionality. At the time of clustering, it is required to reduce dimensionality and redundancy. This section shows the strategies involved in the clustering and summarizes the algorithms.

- Subspace Clustering
  Here the feature matrix is preferred to cluster the objects simultaneously incorporating the data objects as they are maintained in rows. The most common forms of subspace clustering include Two-way clustering, Co-Clustering,bi-clustering, CLIQUE-Clustering in Quest.

3.1 Major Steps of the CLIQUE Algorithm include the following :

- Utilizing Apriori guideline, the subspaces, containing groups, are developed by dividing the information space and identifying the number of concentrates resting within each segmented cell and performing the classification of the subspaces containing bunches.

- Detecting clusters for searching the dense units present in all subspaces of importance along with determining connected dense units in all subspaces of interests

- Producing minimal descriptions for the clusters to find maximal regions that cover a cluster of connected dense units for each cluster along with determining minimal cover for each cluster.

- **Projected Clustering**

  To ensure that the projection into the subspace becomes a tight cluster, it is associated with a subset of a low-dimensional subspace. Projected clustering is recognized with a low-dimensional space sub-set. ORCLUS-Oriented calculation of the expected cluster is an increase in the suggested PROCLUS.

3.2 Major Steps of the PROCLUS Algorithm includes the following:

- Determine a good subset of the piercing set of medoids through a greedy algorithm keeping a small constant.
- Iteratively replace the bad medoids by enhancing the cluster quality in the present set with the new medoids.
- Determining the outliers in the final pass over the data.
- Hybrid Clustering Algorithm
- It intends to produce all subspace clusters along with a number of overlapping points.
- FIRES algorithm is used as a primary approach for the subspace clustering algorithm to generate all the subspace clusters.

3.3 Major Steps of the FIRES Algorithm includes the following

- Computing all the base clusters in the pre-clustering stage.
- Determining the maximal-dimensional subspace clusters.
- Postprocessing steps involve pruning and refinement for enhancing the quality of the mergeable-cluster-set by eliminating the unwanted clusters along with removing the noises thus providing a clean subspace cluster.
- Correlation Clustering

  It is connected in a high dimensional space with the element vector of
- associations between features.
- It can be taken as Biclustering as both share similarity.

3.4 Major Steps of the Correlation Clustering Algorithm include the following

- Selecting each gene in a bicluster specifically through a subset of conditions.
- Selecting each condition through a subset of the genes

3.5 Comparison of the high dimensional algorithms

Algorithm's Name	Advantages	Disadvantages	Future work
CLIQUE	It automatically identifies the subspaces with the maximum dimensionality including high-density clusters.	The accuracy of the method is degraded due to the simplicity of the method.	It can be used in designing of effective clusters to reduce the time complexity along with improving the quality and optimality.
PROCLUS	It is highly suitable for customer classification and trend analysis where partitions of points are essential.	It is suitable for applications where disjoint partitions of the dataset are needed.	Future works include more research to make it applicable to all kind of applications.
FIRES	The largest dimension group approximations from 1D clusters are shown that can be refined to get the real groups.	It is complex to linearlyscale the merge step o f FIRES algorithm.	Future works will be carried out for including the practicality of FIRES to identify intriguing and inevitably important clusters in different data sets of quality expression. .
Correlation Clustering	It finds the set of notable biclusters in a matrix.	There is no guarantee that the reachable remains NP-complete even in the case of planar graphs.	More workneeds to bedone in this type of clustering.

4. Conclusion

The latest development in correspondence and innovation zones takes into account gigantic development in high-dimensional information spaces. This research focuses on issues and actual inadequacies of the existing computations. As the dimensionalities increases, new clustering strategies will emerge. It has been observed that data are extremely sparse in high measurements, along with the extension of separation measures. The primary test for gathering high-dimensional data is still to overcome the "dimensionality scourge." There are various types of continuing ways to handle high-dimensional group data.These techniques need to be compared to produce a better understanding of their strengths and limitations.

5. References

1. Poc, Ángel. "Clustering Algorithms for High-Dimensional Data." (2018).
2. Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In New directions in statistical physics (pp. 273-309). Springer, Berlin, Heidelberg.
3. Babu, B. H., Chandra, N. S., & Gopal, T. V. (2011). Clustering Algorithms For High Dimensional Data–A Survey Of Issues And Existing Approaches. Special Issue of International Journal of Computer Science & Informatics, 2(1), 2.
4. Elankavi, R., Kalaiprasath, R., & Udayakumar, D. R. (2017). A fast clustering algorithm for high-dimensional data. International Journal Of Civil Engineering And Technology (Ijciet), 8(5), 1220-1227.
5. Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional data clustering. Computational Statistics & Data Analysis, 52(1), 502-519.
6. Pavithral, M., & Parvathi, R. M. S. (2017). A Survey on Clustering High Dimensional Data Techniques. International Journal of Applied Engineering Research, 12(11), 2893-2899.
7. Parsons, L., Haque, E., & Liu, H. (2004). Subspace clustering for high dimensional data: a review. Acm Sigkdd Explorations Newsletter, 6(1), 90-105.
8. WIREs Data Mining Knowl Discov 2012, 2: 340–350 doi: 10.1002/widm.1062
9. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. PloS one, 14(1), e0210236.