COMPARING THE TECHNIQUES OF CLUSTER ANALYSIS FOR BIG DATA

Miss. Harshada S. Deshmukh, Prof. P. L. Ramteke

Abstract— **Clustering is an essential data mining and tool for analyzing big data. There are difficulties for applying clustering techniques to big data duo to new challenges that are raised with big data. As Big Data is referring to terabytes of data and clustering algorithms are come with high computational costs, the question is how to cope with this problem and how to deploy clustering techniques to big data and get the results in a reasonable time. Clustering helps to visually analyze the data and also assists in decision making. Clustering is widely used in variety of applications like marketing, insurance, surveillance, fraud detection and scientific discovery to extract useful information.**

In this paper we have discussed some of the current big data mining clustering techniques and also provides a comparison among them.

Index Terms— **Big Data, clustering techniques, Data mining.**

I. INTRODUCTION

The amount of data has been increasing at a faster rate. Since the data has been increasing at a faster rate, it is of great challenge to manage the data. Data Mining is a process of discovering meaningful patterns and rules and to find the relationship among the data. Data Mining is an analytic process with great potential, designed to explore large amounts of data also known as “big data” and search for consistent patterns and systematic relationships between variables, and then to validate the findings by applying the detected patterns to form new subsets of data. It is an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, business intelligence, statistics, high performance computing and database systems. The ultimate goal of data mining is prediction that has the most direct business applications. Data mining software is one of a number of analytical tools for analyzing big data.

In this paper, we introduce the most popular Big data’s clustering techniques. Most state of the art papers found in literature focus on a single category of clustering techniques whereas our goal here is to make a broad and general synthesis concerning the Big Data clustering important techniques.

Clustering is one of the most fundamental technique in big data mining. Clustering is a process of dividing the data elements into groups which are similar to each other. Each group is referred to as a cluster that consists of objects that are similar to one another and dissimilar to objects of another group. It is a technique that recognizes different patterns of data. Good clustering techniques will produce a good or a high quality cluster.

II. TYPES OF CLUSTERING TECHNIQUES

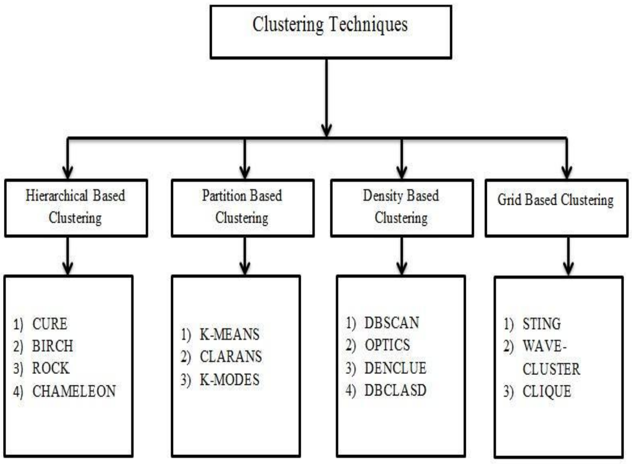

Figure 2.1: Representation of types of clustering techniques

A. *Hierarchical based clustering*

Hierarchical based clustering is also known as connectivity based clustering. It is a method in which hierarchies of clusters are constructed. In this type of clustering, small clusters are merged into a larger one and a large cluster are splinted into smaller clusters. The clusters are constructed by partitioning the instances into a top down or a bottom up approach which can be visualized as a tree like diagram called a “Dendogram” that records the sequence of merges or splits and also shows how the clusters are related. Once the desired numbers of clusters have been formed, the process of splitting or merging will stop. Each cluster nodes consists of child nodes and the node that belongs to the same parent are called as sibling nodes.

Hierarchical clustering is based on two types of algorithms :

Manuscript received Dec, 2015.

Miss. Harshada S. Deshmukh, *Department of Computer Science and Information Technology, HVPM COET, Amravati, India.*
Prof. P.L.Ramteke , *Department of Information Technology, HVPM COET, Amravati, India.*

1) Agglomerative algorithm**:** It is a bottom up approach. It starts by merging each object which are closer to each other or by merging a number of smaller clusters into a larger cluster or until all the objects are merged and a termination condition is met.

Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions). These methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (known as "chaining phenomenon", in particular with single-linkage clustering). In the general case, the complexity is $\mathcal{O}(n^3)$ for agglomerative clustering$\mathcal{O}(2^{n-1})$.

2) Divisive algorithm**:** It is a top down approach. It starts by splitting a larger cluster into smaller clusters until there remain only clusters of one data object and the termination condition is met. For divisive clustering, which makes them too slow for large data sets. For some special cases, optimal efficient methods (of complexity$\mathcal{O}(n^2)$) are known: SLINK for single-linkage and CLINK for complete-linkage clustering

B. Partition Based Clustering

Partition based clustering is a method in which a number of objects are given and the data sets will be partitioned into a number of clusters and each cluster contains similar objects [2]. It generates a specific number of flat and dis-joint clusters and the clusters that are formed will be represented by a centroid or a cluster representative.

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find the $k$cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximative method is Lloyd's algorithm,[8] often actually referred to as "*k-means algorithm*". It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means).

Most k-means-type algorithms require the number of clusters - k- to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters (which is not surprising, as the algorithm optimized cluster centers, not cluster borders).

C. Density based clustering

In this type of clustering, the data objects are separated based on their connectivity, boundary or their region which plays a vital role in finding non-linear shape structure based on the density. This type of clustering helps to separate low dense region (noise data) from high dense region of clusters.

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

The most popular density based clustering method is DBSCAN. In contrast to many newer methods, it features a well-defined cluster model called "density-reachability". Similar to linkage based clustering; it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range queries on the database - and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times. OPTICS is a generalization of DBSCAN that removes the need to choose an appropriate value for the range parameter ε, and produces a hierarchical result related to that of linkage clustering. DeLi-Clu, Density-Link-Clustering combines ideas from single-linkage clustering and OPTICS, eliminating the εparameter entirely and offering performance improvements over OPTICS by using an R-tree index.

The key drawback of DBSCAN and OPTICS is that they expect some kind of density drop to detect cluster borders. Moreover, they cannot detect intrinsic cluster structures which are prevalent in the majority of real life data. A variation of DBSCAN, EnDBSCAN, efficiently detects such kinds of structures. On data sets with, for example, overlapping Gaussian distributions - a common use case in artificial data - the cluster borders produced by these algorithms will often look arbitrary, because the cluster density decreases continuously. On a data set consisting of mixtures of Gaussians, these algorithms are nearly always outperformed by methods such as EM clustering that are able to precisely model this kind of data.

D. Grid based clustering

Grid based clustering is a type of clustering that divides the space into a finite number of cells that are known as grids and all the operations of clustering are applied on these cells [2]. The grids are then combined together to construct a grid like format. The space of the data objects is divided into grids. The main advantage of this approach is its fast processing time, because it goes through the dataset once to compute the statistical values for the grids. The accumulated grid-data make grid-based clustering tech-niques independent of the number of data objects that employ a uniform grid to collect regional statistical data, and then perform the clustering on the grid, instead of the database directly. The performance of a grid-based method depends on the size of the grid, which is usually much less than the size of the database. However, for highly irregular data distributions, using a single uniform grid may not be sufficient to obtain the required cluster-ing quality or fulfill the time requirement. STING (Statistical Information Grid based) and Wave Cluster are examples of grid based clustering. The quality of clustering produced by this method is directly related to the granularity of the bottom most layers, approaching the result of DBSCAN as granularity reaches zero. It explores statistical information stored in grid cells. CLIQUE was the first algorithm proposed for dimension –growth subspace clustering in high dimensional space. Wave Cluster does not require users to give the number of clusters applicable to low dimensional space.

III. COMPARISION OF CLUSTERING ANALYSIS TECHNIQUES

Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods. Further data set can be numeric or categorical. Inherent geometric properties of numeric data can be exploited to naturally define distance function between data points.

A. K-Means Clustering

It is a partition method technique which finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, flat (non-hierarchical) clusters. Stastical method can be used to cluster to assign rank values to the cluster categorical data. Here categorical data have been converted into numeric by assigning rank value. K-Means algorithm organizes objects into k –partitions where each partition represents a cluster. We start out with initial set of means and classify cases based on their distances to their centers. Next, we compute the cluster means again, using the cases that are assigned to the clusters; then, we reclassify all cases based on the new set of means. We keep repeating this step until cluster means don't change between successive steps. Finally, we calculate the means of cluster once again and assign the cases to their permanent clusters.

1) K-Means Algorithm Properties

- There are always K clusters.
- There is always at least one item in each cluster.
- The clusters are non-hierarchical and they do not overlap.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

2) K-Means Algorithm Process

- The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
- For each data point:
- Calculate the distance from the data point to each cluster. If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
- Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
- The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intra cluster distances and cohesion.

B. Hierarchical Clustering

A hierarchical method creates a hierarchical decomposition of the given set of data objects. Here tree of clusters called as dendrograms is built. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent. In hierarchical clustering we assign each item to a cluster such that if we have N items then we have N clusters. Find closest pair of clusters and merge them into single cluster. Compute distance between new cluster and each of old clusters. We have to repeat these steps until all items are clustered into K no. of clusters.

It is of two types:

1) Agglomerative (bottom up)-

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc.It starts by letting each object form its own cluster and iteratively merges cluster into larger and larger clusters, until all the objects are in a single cluster or certain termination condition is satisfied. The single cluster becomes the hierarchy's root. For the merging step, it finds the two clusters that are closest to each other, and combines the two to form one cluster.

2) Divisive (top down)-

A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering but in the opposite direction. This method starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain.

C. DBSCAN Clustering

DBSCAN (Density Based Spatial Clustering of Application with Noise).It grows clusters according to the density of neighborhood objects. It is based on the concept of "density reachibility" and "density connectability", both of which depends upon input parameter-size of epsilon neighborhood e and minimum terms of local distribution of nearest neighbors. Here e parameter controls size of neighborhood and size of clusters. It starts with an arbitrary starting point that has not been visited. The points e-neighbourhood is retreived, and if it contains sufficiently many points, a cluster is started. Otherwise the point is labelled as noise. The number of point parameter impacts detection of outliers. DBSCAN targeting low-dimensional spatial data used DENCLUE algorithm.

D. OPTICS

OPTICS (Ordering Points to Identify Clustering Structure) is a density based method that generates an augmented ordering of the data's clustering structure. It is a generalization of DBSCAN to multiple ranges, effectively replacing the e parameter with a maximum search radius that mostly affects performance. It is an algorithm for finding density based clusters in spatial data which addresses one of DBSCAN"S major weaknesses i.e. of detecting meaningful clusters in data of varying density. It outputs cluster ordering which is a linear list of all objects under analysis and represents the density-based clustering structure of the data. Here parameter epsilon is not necessary and set to maximum value. OPTICS abstracts from DBSCAN by removing this each point is assigned as „core distance", which describes distance to its MinPts point. Both the core-distance and the reachability-distance are undefined if no sufficiently dense cluster w.r.t epsilon parameter is available.

E. STING

STING (STastical INformation Grid)is a grid-based multi resolution clustering technique in which the embedded spatial area of input object is divided into rectangular cells. Statistical information regarding the attributes in each grid cell, such as the mean, maximum, and minimum values are stored as statistical parametersin these rectangular cells. The quality of STING clustering depends on the granularity of the lowest level of grid structure as it uses amultiresolution approach to cluster analysis. Moreover, STING does not consider the spatial relationship between the children and their neighboring cells for construction of a parent cell. As a result, the shapes of the resulting clusters are isothetic, that is, all the cluster boundaries are either horizontal or vertical, and np diagonal boundary is detected. It approaches clustering result of DBSCAN if granularity approaches 0. Using count and cell size information, dense clusters can be identified approximately Using STING.

IV. FUTURE WORK

During the survey, we also find some points that can be further improvement in the future using advanced clustering technique to achieve more efficient accuracy in resultand reduce the time taken for data or information retrieval from large data set.

In the field of crime detection area using clustering techniques, there always remains a scope of improvement in terms of visual, intuitive and investigation techniques that can be developed in an effective way for the detection of crime and social link networks can be developed to link the criminals and study their interrelationships. And the Education, there are large numbers of factors that play an important role in prediction other than the academic which includes non-cognitive factors, to measure and monitor these factors; suitable data mining techniques are required.

V. CONCLUSION

We have studied various clustering techniques which are currently used for analyzing a big data. All these recent techniques are compared on the basis of execution time and cluster quality and their merits and demerits are provided. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, and pattern recognition. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait – often proximity according to some defined distance measure. Data clustering algorithms can be hierarchical or partitioned. Density based clustering is designed for building clusters of arbitrary shapes. It builds clusters automatically i.e. no need to mention the number of clusters and naturally removes outliers. Grid based clustering mainly concentrates on spatial data. EM algorithm provides excellent performance with respect to the cluster quality, excluding for high-dimensional data.

ACKNOWLEDGMENT

I express my sincere gratitude to Resp. Prof. P. L. Ramteke, Head of the Department of Information Technology & for providing their valuable guidance and necessary facilities needed for the successful completion of this seminar throughout.

Lastly, but not least, I thank all my friends and well wishers who were a constant source of inspiration.

REFERENCES

[1] R. Xu and D. Wunsch, ``Survey of clustering algorithms,''IEEE Trans.Neural Netw., vol. 16, no. 3, pp. 645678, May 2005.

[2] H.-S. Park and C.-H. Jun, ``A simple and fast algorithm for K-medoidsclustering,''Expert Syst. Appl., vol. 36, no. 2, pp. 33363341, Mar. 2009.

[3]G.Karypis,E.-H.Han,andV.Kumar,``Chameleon:Hierarchicalclusteringusing dynamic modelling,''IEEE Comput., vol. 32, no. 8, pp. 6875, Aug 1999.

[4] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta,"A Comparative Study of Various Clustering Algorithms in Data Mining,"International Journal of Engineering Reserch and Applications (IJERA),Vol. 2, Issue 3, pp.13791384, 2012.

[5] Arockiam, L., S.S. Baskar, and L. Jeyasimman. 2012. Clustering Techniques in Data Mining.

[6] Garima Sehgal, Dr. Kanwal Garg "Comparison of Various Clustering Algorithms" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3074-3076

[7] Osama Abu Abbas "Comparisons between data clustering algorithms"

[8] Preeti Baser, Dr. Jatinderkumar R. Saini "A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets".

[9] Shafiq Alam, Gillian Dobbie, Patricia Riddle, M. Asif Naeem, "Particle Swarm Optimization Based Hierarchical Agglomerative Clustering", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 64-68.

[10] Kriegel, Hans-Peter; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery **1** (3): 231–240. doi:10.1002/widm.30.

AUTHOR DETAIL:

Miss. Harshada S. Deshmukh is currently pursuing Master in Computer Science and Information Technology at HVPM COET Amravati, Maharashtra (India).

Prof. P. L. Ramteke is currently working as associate professor in Information Technology Department at HVPM COET Amravati, Maharashtra (India).