A Survey on Different Efficient Clustering Techniques used in Web Mining

Mr. Dushyantsinh Rathod[1] Dr. Samrat Khanna[2] Mr. Vijaykumar Gadhavi[3]
[1]Research Scholar [2,3]HOD
[1,2]Department of Computer Engineering
[1]Rai University, Dholka [2]ISTAR College, Vallabh Vidyanagar [3]Aadhiswar College of Engineering

Abstract— Clustering is a process of putting similar data into groups. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. This paper reviews six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DBSCAN clustering, OPTICS, STING.

Key words: Data clustering, K-Means Clustering, Hierarchical Clustering, DBSCAN Clustering, OPTICS, STING

I. INTRODUCTION

Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc. Common approach for all clustering techniques is to find clusters centre that will represent each cluster. Cluster centre will represent with input vector can tell which cluster this vector belong to by measuring a similarity metric between input vector and all cluster centre and determining which cluster is nearest or most similar one

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. Whereas categorical data can be derived from either quantitative or qualitative data where observations are directly observed from counts

II. VARIOUS DATA CLUSTERING TECHNIQUES

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 [2].

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 does'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 intracluster distances and cohesion [6].

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 has 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 [5]

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 [4].

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 [1]. 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 [4].

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 parametre with a maximum search radius that mostly affects performance. MinPts then essentially becomes the minimum cluster size to find. 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 [1].

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 parameters in these rectangular cells. The quality of STING clustering depends on the granularity of the lowest level of grid structure as it uses a multiresolution approach to cluster analysis. Moreover, STING doesnot consider the spatial relationship between the children and their neighbouring 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 [4].

III. CONCLUSION

K-mean algorithm has biggest advantage of clustering large data sets and its performance increases as number of clusters increases. But its use is limited to numeric values. Therefore, Agglomerative and Divisive Hierarchical algorithm was adopted for categorical data, but due to its complexity a new approach for assigning rank value to each categorical attribute using K- means can be used in which categorical data is first converted into numeric by assigning rank.

Hence performance of K- mean algorithm is better than Hierarchical Clustering Algorithm.

Density based methods OPTICS, DBSCAN are designed to find clusters of arbitrary shape whereas partitioning and hierarchical methods are designed to find the spherical shaped clusters.

Density based methods typically consider exclusive clusters only, and do not consider fuzzy clusters. Moreover, density based methods can be extended from full space to subspace clustering.

STING is a query-independent approach since the statistical information exists independently of queries. It is a summary representation of the data in each grid cell, which can be used to facilitate answering a large class of queries, facilitates parallel processing and incremental updating and hence facilitates fast processing.

Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone.

DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to K-Means. Moreover, DBSCAN requires just two parametres and is mostly insensitive to the ordering of points in the database but it cannot cluster data sets well with large difference in densities.

REFERENCES

[1] 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.1379-1384, 2012.

[2] Patnaik, Sovan Kumar, Soumya Sahoo, and Dillip Kumar Swain, "Clustering of Categorical Data by Assigning Rank through Statistical Approach," International Journal of Computer Applications 43.2: 1-3, 2012.

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

[4] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491.

[5] Improved Outcome Software, Agglomerative Hierarchical Clustering Overview. Retrieved from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.htm [Accessed 22/02/2013].

[6] Improved Outcome Software, K-Means Clustering Overview. Retrieved from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm [Accessed 22/02/2013].

