International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Survey of Different Clustering Algorithms in Data Mining

Subodh Shrivastava[1], Brajesh Patel[2]

[1]M.E. Scholar, Department of CSE, SRIT, Jabalpur, India

[2]HOD, Department of CSE, SRIT, Jabalpur, India

Abstract: ***Clustering is the process of collecting similar data in to a group. Clustering is very important in various fields such as machine learning, pattern recognition and statistics. Clustering is a unsupervised learning technique. In this paper following clustering techniques have been discussed- K-Medoid, K-Means Clustering, DBSCAN clustering, IDBSCAN Hierarchical Clustering, OPTICS.***

Keywords: Clustering, DBSCAN, IDBSCAN, OPTIC.

1. Introduction

The process of discovering the knowledge or finding the hidden pattern from a database is known as Data mining. The main goal of data mining is to explore the hidden pattern which was not previously discovered. Various automated decisions can be found out by the Data mining. In data mining we use classification and clustering. Classification is done by supervised learning and clustering is done by unsupervised learning. When training set is provided for learning which act as a example for classes then this process is called supervised learning. When no training set is provided for learning then the process is called as unsupervised learning. In other words we can say that learning from observations and experiments is known as unsupervised learning [1].

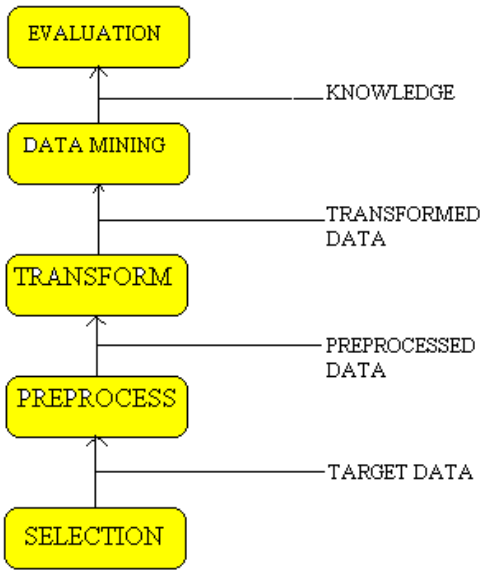

Figure 1: Data Mining Phases

Classification is based on various factors such as kind of database, kind of knowledge, kind of techniques, types of applications. The process in which the objects having similarity are putting in a similar group is called Clustering. It is an approach for finding the similar data which can be used for predicting the information.

2. Data Clustering Methods

2.1 Partitioning Clustering Algorithms

When k clusters are decomposed from set of N objects and a criterion function is optimized then this algorithm is used. With the help of partitioning algorithm the given clustering criteria is minimized by relocating data points again and again until a optimal partition is attained.

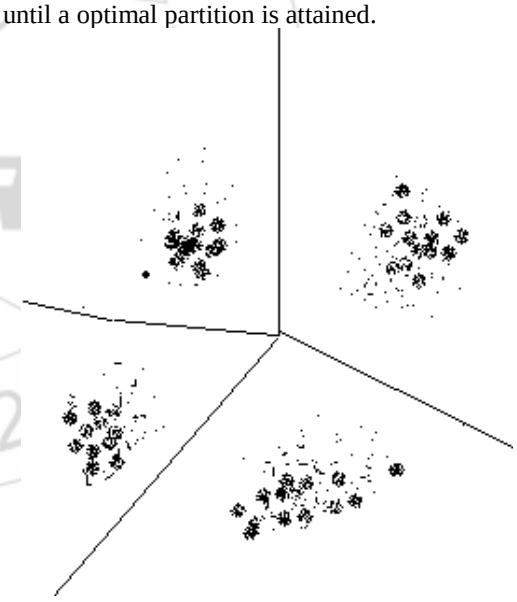

Figure 2: Partitioning Clustering

It constructs a partition of a database in to k clusters, here the number of clusters is defined by the user. There are two types of partitioning algorithms K-means and K-medoid.

2.1.1. K-Means Algorithm

In K-means algorithm first the number of clusters are decided then Initialize the center of the clusters, then assigning all points to closest centroid from the clusters then again compute the centroid and repeat this process until the centroids do not change[2].

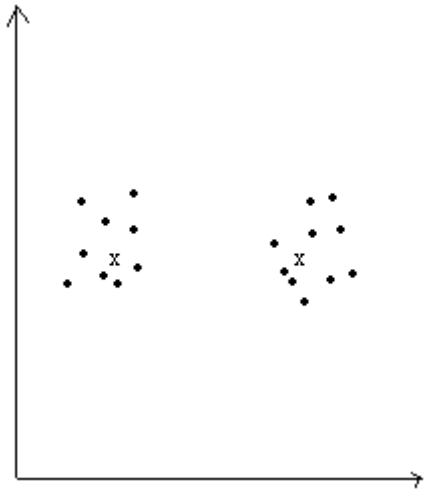

Figure 3: K-Means Algorithm

2.1.2. K-Medoid algorithm

It is also known as PAM-partition around medoids. In this algorithm each cluster is represented by one object which is located near the center. Here we pick actual object to represent cluster instead of taking mean value of the object in the cluster. Then replace again and again one of the medoid by a non medoid and if it improves the total distance of the resulting cluster.

2.2. Hierarchical Algorithms

A tree of clusters is formed by this algorithm. In this algorithm two types of approaches are used top down approach and bottom up approach. In bottom up approach in each step two clusters are merged to form a single cluster and in top down approach each cluster is spitted in to two clusters.

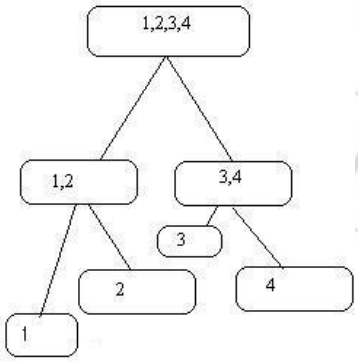

Figure 4: Hierarchical Algorithm

2.3 Density Based Algorithm

This algorithm finds the clusters of data points based on density. In each instance the neighbor of radius given has to contain minimum number of points. Arbitrary shape clusters are formed by the density based algorithm and in single scan it can handle noise very efficiently. DBSCAN, IDBSCAN, OPTICS, CLIQUE.

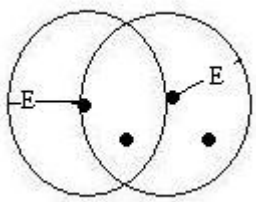

Figure 5: Density Based Algorithm

The points which are in the scope of radius E are defined as core neighborhood. A core point is that point if it has minimum number of specified points. In the above figure the points which are located at the centre is core point. The points which lie on circumference are called border point. The points which are not found within the radius are called outliers or the noise [3].

2.3.1. DBSCAN

In Density Based Spatial Clustering of the Application with Noise maximal set of densely connected points is known as cluster.

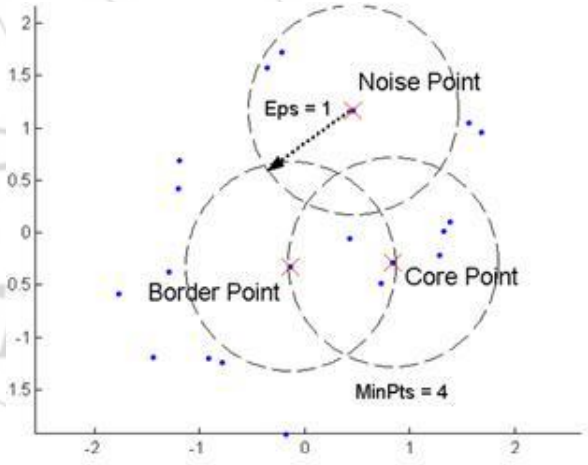

Figure 6: DBSCAN

In DBSCAN algorithm a point p is selected arbitrarily. Then find the points which are reachable from p and which are within the and also satisfied the condition of minimum points. A cluster is formed if point p is core point then. If point p is lie on the circumference or the outer border then no points are density reachable and DBSCAN visits for the next point. This process is repeated until all the points are discovered [4].

2.3.2. IDBSCAN

This algorithm applies marked boundary point to find the data point of an expansion seed. Assume that the core point is P(0,0), the eight marked points may be defined as: A(0, ε), B(0, - ε), C(ε /√2, ε / √2), D(ε, 0), E(ε /√2, - ε / √2}), F(-ε /√2, - ε / √2), G(-ε /√2, ε /√2) , H(-ε, 0),. If P is the core point, and it also satisfies the condition of set density, then the algorithm finds closest point to these eight marked boundary objects marked, and then it sets these points as the expansion seeds. But these seeds may be selected by multiple marked boundary points, the algorithm needs only one instance of input[5].

2.3.3. OPTICS

OPTICS is defined as Ordering Points to Identify Clustering Structure that generates an incremented ordering of data. It is a generalized form of DBSCAN. It replaces the radius with a maximum search radius. MinPts defines the number of points in a cluster size. It is mainly used for spatial data mining. It over comes the problem of DBSCAN algorithm of creating clusters of varying density. In this algorithm the radius of the cluster can be set to maximum value. Therefore this algorithm can make clusters better then the DBSCAN algorithm [6].

2.3.4. CLIQUE

In this algorithm clusters of high density are automatically find. CLIQUE algorithm is a combination of grid based and density based algorithm. Large databases can be clustered by using this algorithm. It partitions the space in to rectangular units and finds the highly dense regions. It uses apriority property to find the required cluster [7].

3. Conclusion

Clustering is the basis for any data analysis. Clustering can be done either by three ways partitioned method, hierarchical method or by density based method. In this survey paper we have define these methods. Partitioned clustering method is fast but it is not fast as hierarchical based method. Spherical clusters are formed by the partitioned method. The clusters made by hierarchical based methods are easily interpretable. But both partitioned and hierarchical methods are not giving better results for large data sets and they are not giving clusters of arbitral shape, only spherical clusters are formed by these methods. Density based methods formed the clusters based on the density of the objects, so clusters of arbitral shape can be formed by these methods which is not possible in the partitioned based and hierarchical based methods.

References

[1] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491. B. Borah and D. K. Bhattacharyya, "An Improved Sampling-Based.

[2] 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].

[3] Le Khac, N.A.; Kechadi, M. "On a Distributed Approach for Density-Based Clustering", *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, On page(s): 283 - 286 Volume: 1, 18-21 Dec. 2011

[4] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density based algorithm for discovering clusters in large spatial databases," in Knowledge Discovery and Data Mining, 1996.

[5] 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.

[6] M. Ankerst and M. M. Breunig and H. P. Kriegel and J. Sander,"OPTICS:Ordering Points to Identify the Clustering Sructure," *Proc. Of ACM SIGMOD Int. Conf. on Management*

[7] Gary Kochenberger, Fred Glover, Bahram Alidaee and Haibo Wang, "Clustering of Microarray data via Clique Partitioning",Journal of Combinatorial Optimization Volume 10, Number 1, 77-92, DOI: 10.1007/s10878-005-1861-1.

Author Profile

Subodh Shrivastava received the B.E.degree in Information Technology from Hitkarini College of Engineering & Technology in 2007. During 2008-2010, he worked as a lecturer in GGCT Jabalpur.