Clustering Techniques and Algorithms of Data Mining – A Review

Krishna Mohan Yadav[1*], Jahanvi Gupta[2], Kartik Jataria[3], Chandra Prabha[4]
[1,2,3]*Student, Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India*
[4]*Assistant Professor, Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India*
*Corresponding author: kmohan770@gmail.com

***Abstract*: Data Mining can be inferred as a result due to the natural evolution of Information Technology. Data mining process is a series of action that ultimately leads to the extraction of information in the form of patterns and records from a large datasets and transforms it into some useful structure for further use. A data mining process is an analysis of data quantities having very large size in order to extract previously unknown patters such as clustered data records, trend, anomalies and dependencies. Data mining becomes a vast area of research in the past few years. This paper describes various clustering based techniques and algorithms along with their advantages and disadvantages. Also, the paper proposes a brief comparative analysis on the performance of clustering algorithms in terms of their complexity, scalability, and robustness.**

***Keywords*: Cluster based mining, Clustering algorithms, Data mining.**

1. Introduction

In software industry, Data Mining has become as a vast research area and it is also most complex and challenging task for developers. Data Mining is defined as the process of fetching the data from the large transactional databases such as stock market, time-series data, spatial databases, multimedia databases, World Wide Web, text databases, medical databases, criminal databases or some specific application oriented databases. Voyage of Knowledge discovery [1][6] is an iterative process with the following steps that is depicted in Fig. 1.

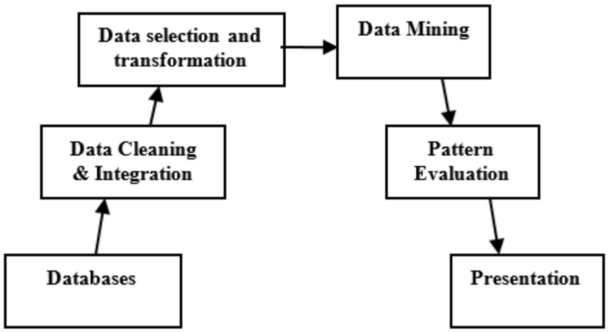
Data selection and transformation
Data Mining
Data Cleaning & Integration
Pattern Evaluation
Databases
Presentation

Fig. 1. Steps of knowledge discovery

- *Data Cleaning:* To remove noise and redundant/ inconsistent data.
- *Data Integration:* Combining of Multiple data sources into one.
- *Data Selection:* Retrieval of required data for analysis from database.
- *Data Transformation:* Data are transformed into forms that are appropriate for mining by implementing aggregation operations.
- *Data Mining:* Application of different optimal methods to extract data patterns.
- *Pattern Evaluation:* Identify the patterns.
- *Knowledge Presentation:* Used to present the mined data to the user.

Association rule mining [15] is also an another mining method by predicting the frequent relations and frequent occurrences of patterns in the large transactional databases such as market basket database, stock market etc.

2. Clustering Techniques and its Algorithms

Clustering [2] is a challenging field of research and developers are developing so many clustering algorithms that can be used for various scientific disciplines. Data Clustering is defined as an exploration process and descriptive data analysis technique which is to analysis the multivariate datasets. Multivariate datasets are the sets which are different in sizes in terms of number of objects and dimensions, and different data types. Data Clustering mainly focuses on large datasets with unknown underlying structure. Cluster analysis is a concept based on human activity like grouping of living and non-living things separately. For example, in Google search, with the help of keywords given by user, related data can be collected or fetched from the databases. Another example [3], in earth observation, clustering can be used to identify the area which is similar in size. Clustering can be widely used in many applications such as image pattern recognition, biology, security, business intelligence etc. Clustering is the process of grouping a set of data according to its similarity. Clustering is also called as data segmentation because it partitions the large datasets into small groups according to its similarity. A cluster is a collection of data objects that are similar to other objects within the same cluster which is called as intra similarity clusters and the objects which are dissimilar forms other clusters is called as inter similarity clusters. Some of the requirements of clustering are scalability, ability to handle different types of attributes; in identifying the clusters of different shape, ability to handle noisy and outliers data's and needs high dimensionality in data.

A. *Partitioning Method*

Combinatorial optimization algorithms [14] are otherwise called as iterative relocation algorithm Fig. 2, which is used to relocate the data points to the nearest clusters, until an optimal partition is attained. There is no guarantee for a globally optimal solution and the convergence is local. An initial partition is a first step in optimization. Local search algorithm can be applied to improve the quality of partitioning as the algorithm provides a foundation for partitioning based clustering methods.

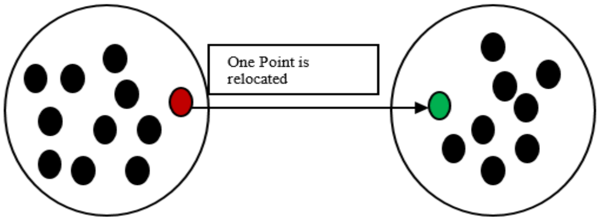

Fig. 2. Iterative relocation algorithm

The methods of partitioning clustering algorithms includes k-Means, k-Medoids, CLARA (Clustering Large Applications) and PAM (Partitioning around Medoids).

k-Means or k-Medoids are reffered as basic iterative algorithm.

1) *k-Means*

The k-Means is a centroid-based technique [1] which can be used where data without defined categories or groups. The algorithm works by taking the input parameter k and partitions a set of n objects into k clusters so that the similarity in intracluster is high when compared with the similarity of intercluster which is low. Based on feature similarity, data points have been clustered. Drawbacks of K Means: It is very sensitive to outliers because an extremely large value object may leads to distortion in the distribution of data. This is because of the use of the square error function.

2) *k-Medoids*

The k-Medoids is otherwise known as representative object based technique [1]. In k-Medoids, instead of taking the mean value of objects as a reference point, take the actual points to represent the clusters by using one representative point per cluster. If the similarity is found in the remaining objects, then that objects are clustered. The main advantage is that the k-Medoids is more robust than the k-Means in the presence of noise and outliers. The main drawback of this algorithm is whenever a point reaches near the center of another cluster; it produces poor outcome because of overlapping in the data points. It uses a number of greedy heuristics schemes for iterative optimization. Its processing is very expensive.

B. *Hierarchical Method*

In Hierarchical clustering [5], the formation of cluster hierarchy or a tree of clusters is called Dendogram. Every node in a hierarchy has child and siblings clusters that are partitioned by the pointscovered by common parent. This type of approach allows the data to explore on different levels of granularity.

Advantages are:

- It has a flexibility regarding the level of granularity.
- It is suitable for all types of attributes.
- Easy to handle the similarity or distance which may be in any form.

Disadvantages are:

- In hierarchical clustering, improvement has not been done because nodes (clusters) cannot be revisited once it has been constructed.
- Termination is not properly done.

Hierarchical Clustering is classified as two categories, namely, Agglomerative nesting and Divisive Analysis.

1) *Agglomerative Nesting*

It is also known as AGNES. Bottom-up approach has been followed fig. 3. In this method, clusters are considered as nodes and it has been constructed like tree structure. The criteria used in this method for clustering the data is min distance, max distance, avg distance, and center distance.

The steps of this method are:

1) Initially all the objects are considered as clusters called as leaf.
2) Recursively, the nodes have been merged according to the maximum similarity between them.
3) Only one cluster has been formed at the end of the process which is called as the root of the tree.

BIRCH:

Balanced Iterative Reducing and Clustering using Hierarchies is a clustering feature process which is used for summarizing the cluster, with the help of CFtree that represents a hierarchy of cluster. It helps to achieve maximum speed and scalability, and is also suitable and very effective for dynamic and incremental clustering of incoming objects.

CHAMELEON:

A Hierarchical Clustering Algorithm [2] follows k-nearest neighbor graph approach which constructs a sparse graph, where vertices can be represented as a data object; the edges were existed between two vertices which contains weight that defines the objects similarity. A graph partitioning algorithm is used to partition the k-nearest neighbor graph which creates so many number of small sub clusters. An agglomerative hierarchical clustering algorithm is used to merge sub clusters repeatedly according to their similarity.

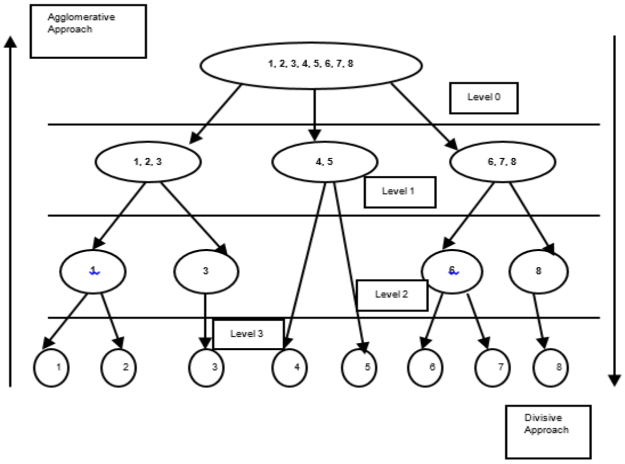

Fig. 3. Representation of AGNES and DIANA

C. *Density Based Method*

Density based algorithms [11] locate the cluster according to the high density regions. It is called as a one-scan algorithms. Density-based connectivity clustering is the first approach that pins density to a training data point. It is having the ability to identify the different shaped clusters and handle noise and outliers very efficiently. Algorithms include OPTICS, DBSCAN, GDBSCAN, and DBCLASD. It is the second approach which pins density to a point in the attribute space and is called Density Functions. DENCLUE is considered as an algorithm for density based method.

1) DBSCAN

The DBSCAN [9] is commonly known as the best density based algorithm. The two parameters of the DBSCAN are Eps and the Minpts. The focus of DBSCAN algorithm is that at least minimum number of points (Minpts) should be present around a point of a given radius of neighborhood points.

2) DENCLUE

DENCLUE [9] is another density based clustering method that deals with density distribution function. A gradient hill-climbing technique is used to find the local maxima of density functions called density attractors, and if kernel density approximation of those local maxima is greater than the noise threshold has been included in the cluster. The drawback [1][2] is that it fails in case of neck type of dataset and it does not work well in case of high dimensionality data.

D. *Grid Based Method*

Grid Density based clustering [8]-[10] mainly focuses on space value that has been surrounded by the data points. Multi resolution grid data structure can be used and it forms clusters by identifying the dense grids. These algorithms requires users to specify a grid size or density threshold, the problem here arise is that how to choose the grid size or density thresholds.

Grid based clustering needs to face the challenges like non-uniformity, locality and dimensionality. To overcome the problem of non-uniformity, adaptive grid based clustering algorithms has been proposed that automatically determines the size of grids based on the data distribution and does not require user to specify any parameter like grid size or density threshold. STING, Wave Cluster, and CLIQUE has been proposed as grid based clustering algorithms.

1) STING

Statistical INformation Grid-based clusterinG is an algorithm that divides the spatial area into rectangular cells and it follows hierarchical clustering. The rectangular cells are formed and categorized as different levels corresponding to different resolution and these cells forms a hierarchical structure. Each cell is partitioned to form a number of cells by combining the high level cells with the next lower level. The calculation for statistical information of each cell has been done and stored which helps to answer queries.

2) CLIQUE

CLIQUE (Clustering in QUEst) is called as a bottom-up subspace clustering algorithm which constructs static grids. Apriori approach has been used to reduce the search space. CLIQUE is a best for density and grid based called subspace clustering algorithm, because it provides high dimensionality and identify the clusters by taking input parameters as density threshold and number of grids. It works well on multidimensional data, but processes a single dimension data at initial step and then grows further upward to the higher level dimensions. The advantage of this algorithm is its quick processing time and independent of the number of data objects. The Disadvantage is that it depends on only the number of cells in each dimension in the quantized space.

3. Literature Survey

K. Chitra and D. Maheswari has proposed the various clustering algorithms along with its advantages and disadvantages The algorithms have been studied in terms of noise, efficiency, scalability, shape of cluster and input data. Supervised and unsupervised learning in mining has been studied. How to measure the similarity in the database, difference in procedure of algorithms, how to find the density threshold and the use of the threshold in clustering has been studied in this paper.

K. Kameshwaran et al. has developed and survey on various clustering algorithms. In this paper, comparison of clustering algorithms with its complexities, advantages and disadvantages has been studied. In this paper, clustering models has been studied and defined. Hierarchical and partitioning based clustering has been proved as connectivity based and centroid based clustering. In this paper, types of clusters have been discussed. One new algorithm in clustering called graph based clustering algorithm has been studied and this algorithm mainly focuses on clustering between the graphs. In this paper, nodes of a graph are clustered and how to implement these clustering by the use of power iteration clustering.

Rui Xu, have developed the survey on clustering algorithms. The procedure for cluster analysis has been discussed with four major steps such as feature selection/extraction, clustering algorithm design/ selection, validation of cluster and interpretation of results. Measure for qualitative features along with its similarity and dissimilarity has been discussed by calculating the distance by implementing the methods such as misskouski distance, Euclidean distance, city block distance, Pearson correlation, point symmetry distance etc. New algorithm called Mixture density based clustering has been proposed that defines the data points in the clusters have been generated by implementing the different probability distribution.

S. Janani and R Tamilselvi has proposed the paper on survey on clustering algorithms. In this paper, Data mining is defined as unsupervised learning task. It provides an overview of clustering types and its algorithms have been studied with its complexities. Clustering with the help of genetic algorithm has been proposed and genetic algorithm is defined as a hypothetical search technique which is used to perform a search

Table 1
Comparison of clustering techniques and algorithms

Key Factors → Approaches ↓		**Complexity for each iteration[17]**	**Performance**	**scalability**	**robustness**
Partitioning Methods	**k-means**	O(nkt)	Not suitable for discovering clusters with non-convex shapes or clusters of very different size	Iterative clustering algorithm is used to improve scalability	Less because of noise and outliers
	k-Medoids	$O(k(n-k)^2)$	Works effectively for small data sets	Parallel k-medoids algorithm based on the Map-Reduce Method	More because medoid is less influenced by outliers
	CLARANS	$O(ks^2+k(n-k))$	Deals well for large data sets	Focusing techniques is used	More effective
Hierarchical Methods	**BIRCH**	O(n)	In a single scan, easily find out the good clustering and improve the quality by few more scans	CF tree which helps to achieve good speed and scalability in large databases and make it effective for incremental and dynamic clustering of incoming objects	Works well with the outliers because the sparse clusters are removed and it is considered as outliers and forms dense clusters into larger ones.
	ROCK	Worst case : $O(n^2+nm_mm_a+n^2logn)$	If the clusters are not spherical in shape, it does not perform well because, the boundary of the cluster has been controlled by the notion of radius or diameter	Random sampling is used for scaling upto large data sets.	Localized approach(that is used to merge into single cluster by finding out the similarity between the objects)are prone to errors.
	Chameleon	$O(n^2)$	Shows greater power`at discovering arbitrarily shaped clusters of high quality.	Sparse graph representation is used to improve the scalabilty	With the help of k nearest neighbor graph, it constructs a sparse graph. This results in the formation of more natural clusters
Density based Methods	**DBSCAN**	$O(n^2)$	Effective at finding arbitrary shaped clusters Not suitable for high dimensional data.	Processing is fast and not efficiently find the clusters if the density of data is changeable.	Works not well in noise
	OPTICS	O(nlogn)	If spatial index is not available, that leads to additional cost while managing the heap.	While finding the cluster border, there will be a decline in density.	Less sensitive to erroneous data
	DENCLUE	O(log|D|)	Allows a brief description of non-spherical shaped cluster in high dimensional data sets.	Processing is much faster.	Very less sensitive to outliers. Detects erroneous data very well.
Grid Based Methods	**STING**	Comp Complexity: O(k), Where k is the no of grid cells in the lower level Time Complexity: O(n) where n is the no of objects	When data is updated, need not to recompute all information again but incremental update is needed. Used for high dimensional spatial data sets.	Answer the queries very efficiently with the help of SQL like languages.	Able to process the dynamically evolving spatial databases
	CLIQUE	Time complexity: $O(c^p+pN)$ it grows exponentially wrt p where p is the highest subspace dimensions selected.	Used for high dimensional data When the algorithm is performing faster, there may be an increase in missing clusters.	Good scalability as the number of dimensions in the data is increased.	Sensitive to noise and insensitive to order of input objects.

in multidirectional way. Individual population is considered as input for genetic algorithm. Various components needed for clustering with genetic algorithm are genetic representation, initial population creation way, function for fitness and genetic operators.

S. Vijayalaksmi, M. Punithavalli has explained about the Clustering Time Series Data Stream - A Literature Survey. In this paper, clustering algorithms that suited for time series data stream has been discussed. The future achievements to be done in time series data stream have been identified.

S. Anitha Elavarasi and J. Akilandeswari has presented the paper on “A Survey On Partition Clustering Algorithms”. His paper gives an overview of various partitioning clustering algorithm and describes about general working behavior, methodologies followed on these approaches and parameters which affects the performance of these algorithms.

Lijuan Zhou and et al. Developed over the study on FP-growth algorithm and proposesed a parallel linked list-based FP-growth algorithm, which is based on Map Reduce programming model, named as the PLFPG (Parallel Linked List-based FPGrowth) algorithm. This algorithm improves the shortcomings of the traditional FP-growth algorithm. First, it describes item-sets space theory, the basic idea of FPGrowth algorithm and the basic components of the Hadoop platform, including HDFS framework and MapReduce programming model. And then, it describes the PLFPG algorithm design ideas. Finally, the algorithm was validated by varying the size of the data set. The results show that the PLFPG algorithm, when compared with the traditional FP-growth algorithm gives a higher operating efficiency and better scalability and extensibility. It can effectively deal with large data sets.

4. Results and Findings

All Clustering techniques and algorithms have been compared in terms of complexity, performance, scalability and robustness. That has been shown in the below Table 1.

5. Conclusion

The main goal of the data mining is to extract the data from the large databases. The main goal of data mining is to separate the data from the large databases and trans-form those data to the form which is suitable for further research study. In this paper, clustering can be studied by various algorithms such as hierarchical, partitioning, density based, and grid based clustering. Through this study, Hierarchical clustering is found to be as connectivity based. Partitioning clustering is called as centroid based. Density based clustering can able to find the high density regions of similar data that form clusters. Grid based clustering defines as the space that can be partitioned into finite number of cells that forms a grid structure. The attributes such as complexity, Scalability, Robust, and performance of all clustering algorithms has been studied. CLIQUE in grid based clustering and Chameleon in hierarchical clustering with the help of sparse graph representations has achieved the good level of scalability. K-Means algorithm in partitioning clustering and OPTICS in density based clustering has achieved more robustness because of less influence of outliers and erroneous data. While studying about the performance, all clustering algorithms are equally achieved the level, but some clustering algorithms are performing well in small datasets whereas some are working well in large datasets. In this criterion, CLARANS in partitioning clustering works well in large data sets as well as very effective in robustness also.

References

[1] K. Chitra and D Maheswari, “A Comparative study of various clustering algorithms in data mining”, International Journal of Computer Science and Mobile Computing, Vol. 6, Issue.8, August 2017, pp. 109-115

[2] K. Kameshwaran and K. Malarvizhi, “Survey on Clustering Techniques in Data Mining”, International Journal of Computer Science and Information Technologies, Vol. 5(2), 2014.

[3] S. Anitha Elavarasi and J. Akilandeswari., “A Survey On Partition Clustering Algorithms”, International Journal of Enterprise Computing and Business Systems (2011).

[4] S.Vijayalaksmi and M Punithavalli., “A Fast Approach to Clustering Datasets using DBSCAN and Applications”, Vol 60– No. 14, pp. 1-7(2012).

[5] P. Berkhin, “A Survey of Clustering Data Mining Techniques” Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press 25- 72(2006).

[6] Mihika Shah and Sindhu Nair., “A Survey of Data Mining Clustering Algorithms”, International Journal of Computer Applications (0975 – 8887) Volume 128 – No.1, October 2015.

[7] Nisha and Puneet Jai Kaur., “A Survey of Clustering Techniques and Algorithms”, IEEE, 2015

[8] Janani and R. Tamilselvi, “A survey of clustering algorithms”, International Journal for Scientific Research & Development (2321-0613), Vol. 5, Issue 10, 2017.

[9] Amandeep Kaur Mann and Navneet Kaur., Survey paper on clustering techniques, International Journal of Science, Engineering and Technology Research, Volume 2, Issue 4, April 2013.

[10] Suman and Pinki Rani., A Survey on STING and CLIQUE Grid Based Clustering Methods, International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017.

[11] Rashi Chauhan, Pooja Batra, Sarika Chaudhary, “A Survey of Density Based Clustering Algorithms” International Journal of Computer Science and Technology, Vol. 5, Issue 2, April - June 2014.

[12] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE,” Survey of Clustering Algorithms”, Published in IEEE transactions on neural networks, vol. 16, no. 3, pp:645-678, May 2015.

[13] Vikas Maral, Sagar Kale, Ketan Balharpure, Sourabh Bhakkad and Pranav Hendre, “Homomorphic Encryption for Secure Data Mining in Cloud”, International Journal of Engineering Science and Computing, Volume 6, issue no 4, April 2016.

[14] Swarndeep Saket J and Dr. Sharnil Pandya, “An Overview of Partitioning Algorithms in Clustering Techniques”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 5, Issue 6, June 2016.

[15] Jeetesh Kumar Jain, Nirupama Tiwari and Manoj Ramaiya, “A Survey: On Association Rule Mining”, International Journal of Engineering Research and Applications, Vol. 3, Issue 1, January-February 2013, pp. 2065-2069.

[16] Hoda Khanali and Babak Vaziri, “A Survey on Clustering Algorithms for Partitioning Method”, International Journal of Computer Applications, Volume 155, No 4, December 2016.

[17] Sabhia Firdaus and Md. Ashraf Uddin, “A Survey on Clustering Algorithms and Complexity Analysis”, IJCSI International Journal of Computer Science Issues, Volume 12, Issue 2, March 2015.