IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

A Comparative Study on Various Clustering Techniques in Data Mining

M. Kasthuri[1*], S. Kanchana[2], R. Hemalatha[3]

[1]Dept. of Computer Science, Bishop Heber College, Tiruchirappalli, Tamil Nadu, India
[2,3]Bishop Heber College, Tiruchirappalli, Tamil Nadu, India

[*]*Corresponding Author: stephenbasilkasthuri@gmail.com*

Available online at: www.ijcseonline.org

Abstract— Clustering is the process of combination of identical objects into same classes. A cluster is a grouping of data objects that are analogous to one another within the same cluster and are disparate to the objects in other clusters. Data clustering can be performed on various areas such as data mining, statistics, machine learning, spatial database, biology and marketing. Machine learning is classified into supervised and unsupervised learning. Clustering is the example of unsupervised learning that has no predefined classes and deals with unknown samples. Cluster analysis can be done with different types of methods includes partitioning methods, hierarchical methods, density based methods, grid based methods and model based methods. Quality of clusters can be determined by the two factors that they are high intra-cluster similarity and low inter-cluster similarity. In this paper, various clustering techniques has been analyzed in data mining in terms of methodology adopted, dataset handled, accuracy, advantages and limitations.

Keywords— Agglomative approach, Clustering, K-means, K-medoid

I. INTRODUCTION

Data mining is a type of sorting technique which is actually used to extract hidden patterns either from large databases, data warehouses or other information repositories [15]. The goals of data mining are fast retrieval of data or information, knowledge Discovery from the databases to reduce the level of complexity and time saving [17].

Data mining also known as KDD (Knowledge Discovery in Databases) which has the following steps: Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Discovery. Cluster analysis can be done by various methods such as Partition based, Hierarchical based, Grid based and Density based algorithms [7]. The discovered knowledge can be applied to decision making, process control, information management, and query processing. Clustering methodologies are very useful in different domains such as prediction analysis, sentiment analysis, forecasting, text categorization, information retrieval [19], Stemming analysis [20], etc.

II. CLUSTERING METHODOLOGIES

A. Partition Method

Partition based techniques divide the object in multiple partitions where single partition describes cluster. Objects within single clusters are of analogous characteristics where an object of different cluster has disparate characteristics in terms of dataset attributes. A distance measure is one of the feature space used to identify similarity or dissimilarity of patterns between data objects. Then it use an iterative relocation techniques that attempts to develop the partitioning by moving objects from one group to another. To achieve global optimality in partition based clustering would require the comprehensive enumeration of all the possible partitions [11]. Instead, most applications implement one of four popular heuristic methods:

1. K-Mean Algorithm : Centroid – Based Technique
2. K-Medoid Algorithm (PAM) : Object – Based technique
3. CLARA (Clustering Large Application): To deal with large data sets
4. CLARANS (Clustering Large Application Based on Randomized Search): It is more capable and scalable than both PAM (Partitioning Around Medoids) and CLARA.

The advantage of the partition based algorithms that they use an iterative way to create the clusters, but the weakness is that the number of clusters has to be determined in advance and only spherical shapes can be strong-minded as clusters. Another major disadvantage of partitioned based algorithm is that whenever the distance between the two points from the centre are close to another cluster, the result becomes poor or misleading due to overlapping of the data points. Table 2.1 shows various clustering methodologies, its advantages and disadvantages of partitioned method.

Table 2.1: Various Methodologies of Partitioned Method

Techniques	Cluster shape	Complexity	Suitable Data set size	Advantages	Disadvantages
K-means	Spherical	O (i k n)	Large	1. K-means is comparatively scalable and efficient processing in large data sets	1. It is difficult to predict the K Value 2. More difficulties in comparing the quality of cluster 3. K-means does not work well with globular clusters 4. It is not suitable for clusters with different sizes and different densities
K-medoid	Arbitrary	O (i k (n-k)2)	Large	1. It is simple to understand and easy to implement 2.K-medoids seems to perform better for large data sets 3. K-medoid is fast and converges in a fixed number of steps 4. Partition Around Medoid (PAM) is less perceptive to outliers than other partitioning algorithms	1. K-mediods is more costly than K-means because of its time complexity 2. Results and total run time depends upon initial partitions
CLARA	Arbitrary	O(ks2+k(n-k))	Sample	1. CLARA deals with larger data sets than PAM	1. The act of CLARA depends upon the size of dataset 2. A partial sample data may result into ambiguous and poor clustering
CLARANS	Arbitrary	O (n^2)	Sample	1. It is easy to handle outliers 2. CLARANS result more effective than PAM and CLARA	1. It does not guarantee to give search to a localized area 2. It uses randomize samples for neighbors 3. It is not much efficient for large datasets

B. Hierarchical Methods

Cluster analysis of hierarchical methods generates the hierarchical decomposition of given data objects. Hierarchical based algorithms can be classified as agglomerative and divisive those are described as bottom up and top down approaches respectively [12]. Agglomerative approach starts with singleton cluster of each separate data object in the dataset and merge them according to the great similar identity adjacently by computing similarity [15]. Divisive approach takes all the data objects as an own single cluster and recursively partition the cluster in order to get one cluster for each data object in the dataset finally. Table 2.2 represents different methodologies of Hierarchical methods.

.

Table 2.2: Different Methodologies of Hierarchical Method

Techniques	Cluster shape	Complexity	Suitable Data set	Advantages	Disadvantages

			size		
BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)	Spherical	O(N)(Time)	Large	1.Efficiently work on given limited amount of main memory and available resources 2.Minimized time required for I/O 3.Applies multiphase clustering technique 4.Produce good quality clustering for single scan and improves the quality of clustering for successive scans	1.Clustering Feature tree can hold only a limited number of entries due to its size 2. BIRCH does not work well for non-spherical shapes because it uses the radius notion to control the boundary of a cluster 3.Works on only numeric dataset
CURE (Clustering Using Representatives)	Arbitrary	O(N2)	Large	1. Overcomes the problem of favoring clusters with spherical shape and similar sizes 2.More robust with respect to outliers 3.High quality clusters in the existence of outliers with complex shapes and different sizes 4.Well clustering quality for large databases 5.One scan of the entire database	1.Sensitive to the user defines parameters-parameter setting does have an significant influence on the results 2.CURE does not handle categorical attributes 3.ignores the information about the aggregate interconnectivity of objects in two different clusters
ROCK (Robust Clustering Using LinKs)	Arbitrary	O(N2)	Large	1.Suitable for categorical attributes	1.Ignores the information about the closeness of two clusters while emphasizing the interconnectivity
CHAMELEON	Graph	O(KN2)	Small	1.Explores dynamic modellings	1.High Time complexity for high dimensional dataset

C. Grid Based Method

The grid based clustering uses a multi resolution grid data structure. It quantizes the space into a limited number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of this approach is its high-speed processing time, which is typically autonomous of the number of data objects, yet reliant on only the number of cells in each dimension in the quantized space. Table 2.3 shows that various methodologies in Grid Based methods. Grid based approach includes:

1. STING (Statistical Information Grid): Explore statistical information stored in grid.
2. Wave Cluster: Objects using a clustering Wavelet Transformation
3. CLIQUE (Clustering In QUEst): Represents a grid and density based approach for clustering in high dimensional data space
4. MAFIA (Merging of Adaptive Intervals Approach to Spatial Data Mining): To handle massive data sets.
5. O-Cluster (Orthogonal partitioning CLUSTERing): Creates a hierarchical grid based clustering model, it creates axis-parallel (orthogonal) partitions in the input attribute space.

Table 2.3: Various Methodologies of Grid Based Method

D. Density Based Method

Techniques	Cluster shape	Complexity	Suitable Data set Size	Advantages	Disadvantages
STING	Arbitrary	O(K)	Large	1. It is a query-independent approach since the statistical information exists independently of queries 2. Query processing using this structure are trivial to parallelize 3. When data is updated, need not recomputed all information in the cell hierarchy	1. All Cluster boundaries are either horizontal or vertical, and no diagonal boundary is selected
MAFIA	Arbitrary	O(cp + p N)	Large	1. MAFIA proposes adaptive grids for fast sub space clustering and introduces scalable parallel frame work 2. It also use shared – nothing architecture to handle massive data sets	1. Gains on higher dimensional data and larger data sets has been observed to be even more dramatic
Wave Cluster	Arbitrary	O(N)	Large	1.Wavelet transformation can automatically result in the removal of outliers 2. The multi resolution property of wavelet transformations can help detect clusters at varying levels of accuracy 3. It can handle any large spatial database efficiently	1. A wavelet transform is only suitable for signal processing techniques, so that decomposes a signal into different frequency is so difficult
O-Cluster	Arbitrary	O(N x d)	Large	1. Good accuracy and scalability 2. It is robust to noise 3. Automatically detects the number of clusters in the data 4.Successfully operate with limited memory resources	1. O-clustering encounter serious scalability and/or accuracy related problems when used on data sets with a large number of records and/or dimensions
CLIQUE	Arbitrary	O(Ck+mk)	Large	1. It automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces 2. It is quite efficient 3. It is insensitive to the order of records in input and does not presume some canonical data distribution	1. The accuracy of the clustering result may be degraded at the expense of simplicity of the method.

Most Density based clustering method is considering distance between objects. These methods can find only spherical shaped clusters and come across difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the view of density [15]. Table 2.4 represents different methodologies in Density method.

1. DEBSCAN: A Density Based clustering method based on Connected Regions with Sufficiently High Density.
2. OPTICS: Ordering Points to identify the Clustering Structure.
3. DENCLUE: Clustering Based on Density Distribution Functions.

Table 2.4: Different Methodologies in Density Based Method

Techniques	Cluster shape	Complexity	Suitable Data set size	Advantages	Disadvantages
DBSCAN	Arbitrary	(N LogN)	large	1. DBSCAN performs efficiently for low dimensional data	1. Sampling data set would affect the density measures 2. The algorithm is not suitable for multiprocessing systems
OPTICS	Arbitrary	O(N Log N)	Large	1.It can discovers the clustering groups with irregular shape, uncertain amount of noise 2. It can discovers high density data included in low density group 3. Final clustering structure is incentive to parameters	1. Expect some kind of density drop to detect cluster borders 2. Less sensitive to outliers
DENCLUE	Arbitrary	O(log |D|)	large	1. It uses solid mathematical foundation and generalizes various clustering methods 2. It is good clustering properties for large data sets	1. Less sensitive to outliers

III. LITERATURE REVIEW OF RECENT CLUSTERING TECHNIQUES

This section is useful to represent literature review about recent clustering techniques in various fields. Table 2.5 shows recent clustering techniques applied in various domains, adapted methodology, data set used, reached accuracy, advantages and Limitations.

Table 2.5: Clustering Techniques adapted in various fields

Year	Authors	Title of the paper	Methodology	Data set	Accuracy	Advantage	Limitations
2018	Pushpalatha K.P., G.Raju	Text Document Retrieval through Clustering using Meaningful Frequent Ordered Word Patterns	Frequent Ordered Word Patterns (FOWPs)	Word Net	85%	1. The rigid nature of Agglomerative Hierarchical Clustering is removed by a greedy approach to select more than one object for clustering in the same iteration using K-nearest Neighbors approach	1. Not able to load the data into RAM fully at a time When the size of the data set is very high
2018	Hemanth Somasekar,Kavya Naveen	Text Categorization and graphical representation using Improved Markov Clustering	Improved Markov Clustering Model (IMCM)	Twenty Newsgroups dataset downloaded from Jason Rennie's page(20 Newsgroups and Reuters-21578)	94%	1. Avoids the overlapped clusters 2. Decreases the CPU time 3. Less memory consumption	1. Not applicable for multilingual document clustering
2018	Dr.Vo Ngoc	English	YULEQ	English	87.85 %	1. It can process	1. Low

	Phu, Dr.Vo Thi Ngoc Tran	sentiment classification using an YULEQ similarity measure and the one dimensional Vectors in a parallel network Environment	similarity coefficient of the clustering technique	testing data set		millions of documents in the shortest time 2. It shortens the execution time in distributed systems 3. It can be applied to other languages	rate of accuracy 2. It takes too much of cost.
2017	Arpit Bansal, MayurSharma and Shalini Goel	Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining	Modified K-mean Clustering and SVM	Cancer data set available in MATLAB	92.68%	1. The modified K-means clustering will vanish off the two major drawbacks, accuracy level and calculation time consumed in clustering dataset	1. This approach is not suitable for large data set
2017	Vairaprakash Gurusamy, S. Kannan J. Regan Prabhu	Mining the Attitude of Social Network Users using K-means Clustering	K-mean, stemming algorithms	Large	87.5%	1. K-means performance is better 2. To save memory space and time	1. K-mean algorithm also increases its time
2016	Shilna S , Navya EK	Heart disease forecasting system using K-mean clustering algorithm with PSO and other data mining methods	PSO (Particle Swarm Optimization) K-Means MAFIA (Maximal Frequent Item set Algorithm)	Cleveland Heart disease Date set	88.20%	1. Reduces main memory needed 2. This prediction handles all missing values and investigates each possibility	1. It consider only a small cluster at a time 2. PSO is a good clustering algorithm, it does not perform well when the dataset is large or complex
2015	K.RajalakshmiDr.S.S.Dhenakaran , N.Roobini	Comparative Analysis of K-Means Algorithm in Disease Prediction	K-means algorithm , Naive Bayes and Decision tree	Patient's data sets	Heart-disease-98.24%, diabetics -78%	1. This system reduces the human effects and cost effective one	1. This paper does not provide clear picture about accuracy and evaluation process
2012	Mohammed M. Abu Tair, Alaa M. El-Halees	Mining Educational Data to Improve Students' Performance: A Case Study	decision tree, rule induction, neural networks, k-nearest neighbor, naïve Bayesian	graduate students data set	-	1.Improves students' performance and also predict current position in academic level.	1. This paper does not provide clear picture about accuracy and evaluation process

IV. CONCLUSION

Clustering is a method of grouping data into different groups, set of objects. The data in each group share similar trends and pattern. This research paper presents detail about various methodologies in clustering technique, which type of data is suitable, its advantages and limitations. Clustering is a significant task in data analysis and data mining applications.

Literature review describes various types of clustering techniques which have been evolved in different domains such as prediction analysis, sentiment analysis, forecasting, text categorization, information retrieval, Stemming analysis [18], etc.

From the literature review it is observed that existing clustering methodologies have its own advantages and limitation. In order to overcome the limitations available in the existing clustering methodologies, if anyone will propose new clustering methodology with hybridize some other techniques like Data structure, Stemming or hybrid clustering, then it is very useful to different domains such as prediction analysis, sentiment analysis, forecasting, text categorization, information retrieval, etc.

REFERENCES

[1] Pushpalatha K.P., and Raju G., "Text Document Retrieval through Clustering using Meaningful Frequent Ordered Word Patterns", In: *International Journal of Applied Engineering Research*, Volume.**13**, pp. **4822-4833**, ISSN: 0973-4562, **2018**.

[2] Hemanth Somasekar and Kavya Naveen, "Text Categorization and graphical representation using Improved Markov Clustering", In: *International Journal of Intelligent Engineering and Systems*,Volume.**11**, pp.**107-116**, March **2018**.

[3] Dr.Vo Ngoc Phu and Dr.Vo Thi Ngoc Tran, "English Sentiment Classification Using an YULEQ Similarity Measure and the One dimensional Vectors in a Parallel Network Environment", In: *Journal of Theoretical and Applied Information Technology*, Volume.**96**, pp.**3356-3382**, ISSN: 1992-8645, June **2018**.

[4] Arpit Bansal, Mayur Sharma, Shalini Goel, "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining", In: *International Journal of Computer Applications*, Volume.**157**, pp.**35- 40**, January **2017**.

[5] Vairaprakash Gurusamy S., Kannan, J. and Regan Prabhu, "Mining the Attitude of Social Network Users using K-means", In: *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume.**7**, Issue.**5,** pp.**226-230**, ISSN: 2277-128X, May **2017**.

[6] Suman, Pinki Rani, "A Survey on STING and CLIQUE Grid Based Clustering Methods", In: *International Journal of Advanced Research in Computer Science*, Volume.**8**, pp.**1510-1512**, ISSN: 0976-5697, May-June **2017**.

[7] Chitra K., and Maheswari D., "A Comparative Study of Various Clustering Algorithms in Data Mining", In: *International Journal of Computer Science and Mobile Computing*, Volume.**6**, Issue.**8**, pp.**109-115**, ISSN: 2320–088X, August **2017**.

[8] Sajana T., Sheela Rani C.M., and Narayana K.V, "A Survey on Clustering Techniques for Big Data Mining", In: *Indian Journal of Science and Technology*, Volume.**9**, pp.**1-12**, ISSN: 0974-6846, January **2016**.

[9] Shilna S. and Navya EK, "Heart Disease Forecasting System using K-mean Clustering Algorithm with PSO and other Data Mining Methods", In: *International Journal On Engineering Technology and Sciences*, Volume.**3**, Issue.**4**, pp.**51-55**, ISSN(P): 2349-3968, ISSN (O): 2349-3976, April **2016**.

[10] Kiranjeet Kaur and Lalit Mann Singh, "Heart Disease Prediction System Using ANOVA, PCA And SVM Classification", In: *International Journal of Advance Research, Ideas and Innovations in Technology*, Volume.**2**, Issue.**3**, pp.**1-6**, ISSN: 2454-132X, **2016**.

[11] Swarndeep Saket J. and Sharnil Pandya, "An Overview of Partitioning Algorithms in Clustering Techniques", In: *International Journal of Advanced Research in Computer Engineering & Technology*, Volume.**5**, Issue.**6**, pp.**1943-1946**, ISSN: 2278–1323, June **2016**.

[12] Abdullah Z. and Hamdan A. R, "Hierarchical Clustering Algorithms in Data Mining", In: *International Journal of Computer and Information Engineering*, Volume.**9**, pp.**2201- 2206**, **2015**.

[13] Rajalakshmi K, and Dhenakaran, Roobini N, "Comparative Analysis of K-Means Algorithm in Disease Prediction", In: *International Journal of Science, Engineering and Technology Research* (IJSETR), Volume.**4**, Issue.**7**, pp.**2697-2699**, ISSN: 2278-7798, July 2015.

[14] Abdullah Z, Hamdan A. R, "Hierarchical Clustering Algorithms in Data Mining", In: *International Journal of Computer and Information Engineering*, Volume.**9**, pp.**2201- 2206**, **2015**.

[15] Jiawei Han and Micheline Kamber, "*Data Mining Concepts and Techniques*", Morgan Kaufmann, Third Edition, **2012**.

[16] Mohammed M. Abu Tair, Alaa M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study", In: *International Journal of Information and Communication Technology Research*, Volume.**2**, pp.**140-146**, February **2012**.

[17] M. Kasthuri, Dr. S. Britto Ramesh Kumar, "A Framework for Language Independent Stemmer Using Dynamic Programming", In: *International Journal of Applied Engineering Research* (IJAER), Print ISSN 0973-4562, Volume.**10**, Number.**18**, pp.**39000-39004** Online ISSN1087-1090, **2015**.

[18] M. Kasthuri, Dr. S. Britto Ramesh Kumar, "PLIS: Proposed Language Independent Stemmer for Information Retrieval Systems Using Dynamic Programming", In: *2016 World Congress on Computing and Communication Technologies*, ISBN: 978-1-5090-5573-9, pp.**132-135**, IEEE, **2016**.

[19] M. Kasthuri and Dr. S. Britto Ramesh Kumar, "Multilingual Phonetic Based Stem Generation", In: *Proceedings of the Second International Conference on Emerging Research in Computing, Information Communication and Applications* (ERCICA-2014), ELSEVIER Science :: Technology India, ISBN:9789351072 607, Volume.**1**, pp.**437-442**, 01-02 August, **2014**.

[20] Dr.M.Kasthuri and Dr.S.Britto Ramesh Kumar, "PLIS: Proposed Language Independent Stemmer Performance Evaluation", In: *International Journal of Advanced Research in Computer Science & Technology* (IJARCST 2017), Volume.**5**, Issue.**4**, ISSN: 2347-8446 (Online), ISSN:2347-9817 (Print), Oct-Dec, **2017**.

Authors Profile

Dr. M.Kasthuri is working as an Assistant Professor in the Department of Computer Applications, Bishop Heber College, Tiruchirappalli, Tamil Nadu, India. She had completed her Doctorate of Philosophy in Computer Science in June 2017 at Bharathidasan University, Tiruchirappalli. She has published a number of National and International level research papers related to Web Mining and Stemming concepts. She has completed UGC sponsored Minor Research Project entitled as Language Independent Stemmer.

S. Kanchana Subramanian is a student of Master of Computer Applications at Bishop Heber College in Tiruchirappalli, She completed her Under Graduate of Bachelor of Computer Applications at Bishop Heber College, Tiruchirappalli in 2016. She secured Gold Medal for Computer Applications in University Rank Examinations 2016 conducted by Bharadhidasan University, Tiruchirappalli, Tamil Nadu, India.

R. Hemalatha R completed her Bachelor degree in computer applications at Bishop Heber College, Trichirappalli in 2016 and currently she studying her master degree in computer applications at Bishop Heber College, Tirchirappalli. She got department second in her under graduate program of computer applications.

