International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 5

A Research on Different Clustering Algorithms and Techniques

M. Pavithra[1], P. Nandhini[2], R. Suganya[1]
[1]Assistant Professor, [2]PG Scholar
Department of C.S.E, Jansons Institute of Technology,
Coimbatore, Karumathampatti, Tamil Nadu India

ABSTRACT
Learning is the process of generating useful information from a huge volume of data. Learning can be classified as supervised learning and unsupervised learning. Clustering is a kind of unsupervised learning. Clustering is also one of the data mining methods. In all clustering algorithms, the goal is to minimize intracluster distances, and to maximize intercluster distances. Whatever a clustering algorithm provides a better performance, it has the more successful to achieve this goal [2]. Nowadays, although many research done in the field of clustering algorithms, these algorithms have the challenges such as processing time, scalability, accuracy, etc. Comparing various methods of the clustering, the contributions of the recent researches focused on solving the clustering challenges of the partition method [3]. In this paper, the partitioning clustering method is introduced, the procedure of the clustering algorithms is described, and finally the new improved methods and the proposed solutions to solve these challenges are explained [4].The clustering algorithms are categorized based upon different research phenomenon. Varieties of algorithms have recently occurred and were effectively applied to real-life data mining problems. This survey mainly focuses on partition based clustering algorithms namely k-Means, k-Medoids and Fuzzy c-Means In particular; they applied mostly in medical data sets. The importance of the survey is to explore the various applications in different domains [5].

Keywords: *Clustering, Supervised Learning, Unsupervised Learning, Data Mining*

1. INTRODUCTION

Since 1990s, the notion of data mining, usually seen as the process of "mining" the data, has emerged in many environments, from the academic field to the business or medical activities, in particular. As a research area with not such a long history, and thus not exceeding the stage of adolescence" yet, data mining is still disputed by some scientific fields [1]. In this sense, data mining means at various references are as follows: •

- Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories [2].
- Data mining is a process that uses algorithms to discover predictive patterns in data sets. "Automated data analysis" applies models to data to predict behavior, assess risk, determine associations, or do other types of analysis [3].

Actually, when data mining methods to solve concrete problems are used, in mind their typology is created, which can be synthetically summarized in two broad categories, predictive methods and descriptive methods, already referred to as the objectives of data mining. Clustering is the type of descriptive methods [1].

Clustering is a process of grouping objects with similar properties. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity. Clustering is an unsupervised learning i.e. it learns by observation rather than examples. There are no predefined class label exists

for the data points. Cluster analysis is used in a number of applications such as data analysis, image processing, market analysis etc. Clustering helps in gaining, overall distribution of patterns and correlation among data objects [1]. This paper describes about the general working behavior, the methodologies to be followed and the parameters which affects the performance of the partition clustering algorithms. This paper is organized as follows; section 2 gives an overview of different clustering algorithms. In section 3 various partition clustering algorithms, the methodology applied on these algorithms and the parameter which has the impact on the efficiency of these algorithms are described. Finally in section 4 the conclusions are provided [7].

In general the clustering algorithms can be classified into two categories, hard and soft (fuzzy) clustering [4]. Fuzzy clustering is a widely applied method for obtaining fuzzy models from data. It has been applied successfully in various fields including geographical surveying, finance or marketing. In classical cluster analysis each datum must be assigned to exactly one cluster. Fuzzy cluster analysis relaxes this requirement by allowing gradual memberships, thus offering the opportunity to deal with data that belong to more than one cluster at the same time [5], and the Boolean-like nature of assignment relaxed by assigning membership grades that assume values in the unit interval and quantify a strength of belongingness of a data point to the individual cluster [6].

All data mining tasks can be categorized in to two types: supervised tasks and unsupervised tasks. Supervised tasks have datasets that contain both the explanatory variables and the dependent variables, and the objective is to discover the associations between the explanatory variables and the dependent variables [9]. On the other hand, unsupervised mining tasks have datasets that contain only the explanatory variables with the objective to explore and generate postulates about the buried structures of the data. Clustering is any of the most common untested data mining methods that explore the hidden structures embedded in a dataset. Clustering has been effectively applied in various engineering and scientific disciplines such as psychology, biology, medical dataset clustering, computer vision, communications, and remote sensing [11]. Cluster analysis organizes data (a collection of patterns, each design could be a direction measurements) by abstracting and an underlying structure. The combination is done such that patterns inside a group (cluster) are more related to each other than patterns belonging to different groups. Therefore, group of data using cluster analysis employs some dissimilarity measure among the set of patterns [10].

CLUSTERING METHODS

In this paper, various methods of clustering are divided into hierarchical clustering, density-based clustering, grid-based clustering, incremental clustering, and partitional clustering, and they are presented.

Hierarchical clustering

A hierarchical algorithm yields a dendrogram representing the nested grouping of patterns and similarity levels at which groupings change [7], and this algorithms are divided into the two categories divisible algorithms and agglomerative algorithms [8] But A major weakness of agglomerative clustering methods is that they do not scale well [9], also hierarchical clustering suffer from the fact that once a step (merge or split) is done, it can never be undone. This rigidity is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices [2]. Some of the interesting studies in this direction are Chameleon [10], Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) [11] , and so on. Hierarchical clustering algorithm groups data objects to form a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, centroid or wards method. In divisive approach all data points are considered as a single cluster and they are splited into number of clusters based on certain criteria, and this is called as top down approach. Examples for this algorithms are LEGCLUST [3], BRICH [2] (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives) [2], and Chemeleon [1].

Density-based clustering

This method clusters based on a local criterion such as density-connected points. The major features of the

clustering include the abilities to discover clusters of arbitrary shape and handle noise. The algorithm requires just one scan through the database. However, density parameters are needed for the termination condition [9]. Density based algorithm continue to grow the given cluster as long as the density in the neighborhood exceeds certain threshold [1]. This algorithm is suitable for handling noise in the dataset.

The following points are enumerated as the features of this algorithm.

1. Handles clusters of arbitrary shape
2. Handle noise
3. Needs only one scan of the input dataset?
4. Needs density parameters to be initialized. DBSCAN, DENCLUE and OPTICS [1] are examples for this algorithm.

Spectral Clustering

Spectral clustering refers to a class of techniques which relies on the Eigen structure of a similarity matrix. Clusters are formed by partition data points using the similarity matrix. Any spectral clustering algorithm will have three main stages [4]. They are

1. **Preprocessing:** Deals with the construction of similarity matrix.
2. **Spectral Mapping**: Deals with the construction of eigen vectors for the similarity matrix
3. **Post Processing:** Deals with the grouping data points

The following are advantages of Spectral clustering algorithm:

1. Strong assumptions on cluster shape are not made.
2. Simple to implement.
3. Objective does not consider local optima.
4. Statistically consistent.
5. Works faster.

Partition Clustering

Partition clustering algorithm splits the data points into k partition, where each partition represents a cluster [4]. The partition is done based on certain objective function. One such criterion functions is minimizing square error criterion which is computed as,

$$E = \sum \sum || p - mi ||^2$$

where p is the point in a cluster and mi is the mean of the cluster. The cluster should exhibit two properties, they are (1) each group must contain at least one object (2) each object must belong to exactly one group. The main drawback of this algorithm [3] is whenever a point is close to the center of another cluster, it gives poor result due to overlapping of data points.

Partition clustering algorithms

K-means:

It starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers until a convergence criterion is met [4]. The method is relatively scalable and efficient in processing large data sets [2], its time and space complexity is relatively small, and it is an order-independent algorithm [8]. But the method often terminates at a local optimum, and is not suitable for discovering clusters with no convex shapes or clusters of very different size [2]. Moreover, an ambiguity is about the best direction for initial partition, updating the partition, adjusting the number of clusters, and the stopping criterion [8]. A major problem with this algorithm is that it is sensitive to noise and outliers [9].

K-medoid/PAM:

PAM was one of the first k-medoids algorithms introduced [2]. The algorithm uses the most centrally located object in a cluster, the medoid, instead of the mean. Then, PAM starts from an initial set of medoids, and it iteratively replaces one of the medoids by one of the nonmedoids if it improves the total distance of the resulting clustering [9]. This algorithm works effectively for small data sets, but does not scale well for large datasets [2].

CLARA:

Instead of taking the whole set of data into consideration, a small portion of the actual data is chosen as a representative of the data. Medoids are then chosen from this sample using PAM. CLARA draws multiple samples of the data set, applies PAM on each sample, and returns its best clustering as the output. As expected, CLARA can deal with larger data sets than PAM [2].

CLARANS:

It draws a sample with some randomness in each step of the search. Conceptually, the clustering process can be viewed as a search through a graph. At each step, PAM examines all of the neighbors of the current node in its search for a minimum cost solution. The current node is then replaced by the neighbor with the

largest descent in costs [2]. The algorithm also enables the detection of outliers [9].

FCM:

Fuzzy c-means algorithm is most widely used [4], and an extension of the classical and the crisp k-means clustering method in fuzzy set domain So that it is widely studied and applied in pattern recognition, image segmentation and image clustering, data mining, wireless sensor network, and so on [5]. The objects of FCM can belong to more than one cluster, as well as a membership grade is associated with each of the objects indicating the degree to which objects belong to the different clusters [6]. Moreover, FCM has better performance in many clustering issues, but its computational complexity is higher than K-means, and FCM is sensitive to noise and outliers.

Grid-based clustering

Grid-based clustering quantizes the pattern space into a finite number of cells. These algorithms typically involve fast processing time since they are dependent only on the number of cells in each dimension of the quantized space and are typically independent of the number of actual data objects [9]. The clustering approach uses a multi resolution grid data structure. The main advantage of the approach is its fast processing time [2].

Incremental clustering

The algorithms of this clustering work on large data sets, where it is possible to accommodate the entire data set in the main memory. The space requirements of the incremental algorithm are very small, necessary only for the centroids of the clusters [7]. Typically, these algorithms are non-iterative and therefore their time requirements are also small. If iterations even are introduce into the incremental-clustering algorithm, computational complexity and corresponding time requirements do not increase significantly [5]. Most incremental algorithms do not satisfy one of the most important characteristics of a clustering process: order-independence. The algorithms are very sensitive to the order of samples, and for different orders they generate totally different partitions [8].

COMPARISON OF CLUSTERING METHODS

According to the above, all algorithms are designed to minimize intra cluster distances, and to maximize inter cluster distances. In Table 1, the clustering methods are briefly described and evaluated.

Table 1. Comparison of clustering methods

Features	Hierarchical clustering	Density-based clustering	Grid-based clustering	Incremental clustering	Partitional clustering
Data size	Small to medium	Small to medium	Small to medium	Large	Small to medium
Data shape	Spherical	Complex and non-spherical	Network of pattern space	Unlimited	Spherical
Based on	Distance, density and continuity	Density	Network structure	Big data	Distance
Advantages	- Decreasing calculations and costs	- Clusters of arbitrary data shape - Suitable for filtering outliers and noisy data - Need to once scan the database - Determine the number of clusters simultaneously with clustering	- Rapid creation of models - A multiresolution grid data structure	- Save data to secondary memory, and transfer to the main memory only once for clustering. - Non-repetitive - low time complexity - Suitable for very large data sets	- Low computational complexity
Disadvantages	- Nonscalable - Inability to reform a wrong decision	- Need to condition the termination	- Low accuracy	- Not satisfy a order-independence	- Need to set the number of clusters - Need to stopping criterion

IMPROVED PARTITION ALGORITHMS

Multi-center Fuzzy C-means algorithm based on Transitive Closure and Spectral Clustering (MFCMTCSC)

It uses the multi-center initialization method to solve sensitive problems to initialize for FCM algorithm, and applies non-traditional curved clusters. To ensure the extraction of spectral features, Floyd algorithm provides a similarity Matrix used block symmetric [8]. On the other the problem of clustering samples is changed into a problem of merging sub clusters, thus the computational load is low, and has strong robustness [9].

Robust clustering approach

It is based on the maximum likelihood principle, and focuses on maximizing the objective function. The approach also extends the Least Trimmed Squares

approach to fuzzy clustering toward a more general methodology [3]. Moreover, it discards a fixed fraction of data. The fixed trimming level controls the number of observations to be discarded in a different way from other methods that are based on fixing a noise distance [6]. This approach also considers an eigenvalue ratio constraint that makes it a mathematically well-defined problem and serves to control the allowed differences among cluster scatters.

FCM-RSVM

It improves the performance of FCM clustering algorithm by combining it with relaxed constraints support vector machine (RSVM) [3] so that first fuzzy c-means partitions data into appropriate clusters. Then, the samples with high membership values in each cluster are selected for training a multi-class RSVM classifier. Finally, the class labels of the remaining data points are predicted by the latter classifier [8].

Relative Entropy Fuzzy C-Means (REFCM)

It adds the relative entropy to FCM‟s objective function as a regularization function. Membership functions in FCM have probabilistic interpretation, and the other the relative entropy is a non-negative and convex function [4]. Therefore, relative entropy is used REFCM. This algorithm minimizes the within clusters distances and meanwhile maximizes the dissimilarity between clusters, and it has the ability to detect noise points and to allocation reasonable membership degrees to observations [6].

Min Max K-Means

It overcomes the initialization problem of k-means by altering its objective such that first the initial centers are methodically picked and second the cluster sizes are balanced [2]. The method applying k-means to minimize the sum of the intra-cluster variances is the appropriate clustering approach, but MinMax K-Means is nonlinearly not designed to separable clusters can be detected in the data [5].

Interval Type-2 Credibilistic Clustering (IT2CC)

It considers both degrees of membership and non-membership in accounts. On the other hand, the method applies the concept of credibility theory to design a new objective function which simultaneously includes compactness within clusters and separation of them. The credibility concept is utilized to integrate degrees of membership and non-membership [8].

COMPARISON OF IMPROVED PARTITION ALGORITHMS

Due to the above subjects, each of these algorithms has proposed the solutions. However, the improved methods have advantages and disadvantages briefly mentioned in Table 2.

According to Table 2, MFCM-TCSC algorithm through the multi-center based on transitive closure and spectral clustering, CGS algorithm through the deterministic selection rules and recursively solving KHM, SC-FCM algorithm through the multidimensional and self-renewal properties of stem cells, MinMax K-Means algorithm through minimizing the maximum intra-cluster variance instead of the sum of the intra-cluster variances, and BFCM, BGK, and BICS algorithms through the bias-correction approach solve the sensitive to initialize. Furthermore, F-TCLUST algorithm through discarding a fixed fraction of data, REFCM algorithm through fuzzy clustering features and relative entropy, and both IT2CC and Multi-central general type-2 fuzzy clustering algorithms through the type-2 fuzzy clustering solve the sensitive to noise and outliers [9].

FCM-RSVM algorithm uses relaxed constraints of support vector machine to assign low membership values of data points in clusters. Here, it should be noted that these algorithms have disadvantages including: MFCM-TCSC algorithm may cause redundancy features [10]. F-TCLUST algorithm does not provide an evaluation of performance of classification curves. FCMRSVM algorithm almost has the sensitivity to over-fitting because it uses RSVM algorithm. CGS algorithm has a large and great candidate group to replace centers in large data sets thus it may not suitable for large instances [11]. SC-FCM algorithm due to the use of SCA optimization algorithm needs relatively large memory. MinMax K-Means algorithm does not use the kernel-based clustering. BFCM, BGK, and BICS algorithms have a high number of iterations despite the use of the bias-correction approach [12]. Because the type-2 fuzzy approach has high computational complexity, it affects the high computational time, and the high iterations of IT2CC and Multi-central general type-2 fuzzy clustering algorithms.

APPLICATIONS OF CLUSTERING ALGORITHMS

Clustering plays a vital role in data analysis. The kmeans algorithm is a simple algorithm that has been adapted to many problem domains. It can be seen that the k-Means algorithm is a blameless candidate for extension to work with fuzzy feature vectors [4]. A large number of clustering algorithms must remain developed in a variety of domains for different types of applications. None of these algorithms is suitable for all types of applications. This survey work is carried out to analyze the applications of partition based clustering algorithms, done by various researchers in different domains [5]. The clustering algorithms has been applied to more applications including improving the efficiency of information retrieval system and simulation of special Medical Cluster applications and some other areas. We proposed a new framework to improve the 'web sessions' in their paper "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm". In this research work, initially a modified k-means algorithm is used to cluster the user sessions [7]. The refined initial starting condition allows the iterative algorithm to converge to a "better" local minimum. Then they have proposed a GA (Genetic Algorithm) based refinement algorithm to improve the cluster quality [6].

Discovering (Uterus disease diagnosis) Interesting patterns (required level of instructions with complete form) in the initial Cluster. Discovering overlapping cluster by the reassignment of medical dataset under disease level [2]. Update the score cache level of the information finding, retrieval phase of the Medical applications with different types of datasets. This article describes the properties of k-means algorithm among the other algorithms and they are summarized below for quick absorbency of the researchers and developers.

- Dependent on starting point and number of clusters can require very little iteration.
- Limits of clusters are well defined without overlapping.
- The results of k Means are reliant on starting Centroid locations.
- Frequently different optimal results can be produced using different starting centroids.
- Creates large clusters early in process; clusters formed are dependent on order of input data.
- Execution efficiency results depend upon how many clusters are chosen.
- Adaptable for isolating spherical or poorly separated Clusters.

CONCLUSION

Data mining includes techniques in various fields to analyze the data. Many algorithms apply different analyzes of data in this field. In this paper, after reviewing clustering methods of data mining, a number of these algorithms are presented as a whole and an independent of the algorithm, and their differences are studied. The discussed methods were an introduction to the concepts and the researches which indicated available algorithms by different functions in any fields [3]. In the following, the new improved algorithms, and the proposed solutions to solve the challenges of the partition algorithms were described. Finally, each clustering algorithm is not generally considered the best algorithm to solve all problems, and the algorithms designed for certain assumptions are usually assigned to special applications. Considering the importance of partitioned clustering in data mining, and its being widely in recent years, clustering algorithms have become into a field of active and dynamic research [7].

Therefore, improving the partition clustering algorithms such as K-means and FCM could be an interesting issue for future research.The paper describes different methodologies and parameters associated with partition clustering algorithms. The drawback of k-means algorithm is to find the optimal k value and initial centroid for each cluster [9]. This is overcome by applying the concepts such as genetic algorithm, simulated annealing, harmony search techniques and ant colony optimization.The choice of clustering algorithm depends on both the type of data available and on the particular purpose and chosen application []. The partitioned algorithms are work well for finding spherical shaped clusters in the given input as Medical dataset. This article discusses the various application areas of partition based clustering algorithms like k-Means, k-Medoids, Fuzzy C-Means [10]. The k-Means algorithm is very consistent when compared and analyzed with the other two algorithms. Further, it stamps its superiority in terms of its lesser execution time. From this survey, it is identified the applications of innovative and special approaches of clustering algorithms principally for medical domain [4]. From the various applications by several researchers, particularly, the performance of k-Means algorithm is well suited. Most of the researchers are

finding that the k- means algorithm is more suitable than other algorithms in their field [6].

REFERENCES

1. Ester, M., Kriegel, H., Sander, J. and Xu, X, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", 2010.
2. Ghosh, Soumi, and Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", IJACSA, Vol.4, No.4, 2013.
3. H. Ming-Chuan, "An efficient fuzzy c-means clustering algorithm", In Proceedings of IEEE International Conference on Data Mining, ICDM-2001, ISBN. 0-7695- 1119-8, pp. 225-232, 2011.
4. Kanungo, T, Mount, D. M, Netanyahu, N. S., Piatko, C. D, Silverman, R. & Wu, A. Y, "A local search approximation algorithm for k-means clustering", 18th Annual ACM Symposium on Computational Geometry (SoCG'02), pp. 10-18, 2012.
5. KSravya and S. VaseemAkram, "Medical Image by using the Pillar K-means Algorithm", International Journal of Advanced Engineering Technologies, Vol. 1, Issue 1, 2013.
6. Masulli, Francesco and Andrea Schenone, "A fuzzy clustering based segmentation system as support to diagnosis in medical imaging", Artificial Intelligence in Medicine Vol. 16, Issue 2, pp. 129-147, 2009.
7. Nidal M, Christoph. F, "K-medoid-style Clustering Algorithms for Supervised Summary Generation", in Proc. of the International Conference on Machine Learning; Models, Technologies and Applications (MLMTA'04), pp. 932-938, 2014.
8. SrinivasaPerumal. R, R. Sujatha, "Analysis of Colon Cancer Dataset using K-Means based Algorithms & See5 Oriental Algorithms", IJCST, Vol. 2, No. 4, pp. 482-485, 2011.
9. [9]. Sujatha, N and K. Iyakutty, "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm", European Journal of Scientific Research, pp. 478-490, 2010
10. Velmurugan. T, "Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points", International Journal of Computer Technology & Applications, Vol. 3, Issue 5, pp. 1758-1764, 2012.
11. Wei Zhong, GulsahAltun, Robert Harrison, Phang C. Tai, and Yi Pan, "Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property", IEEE Transactions on NanoBioscience, Vol. 4, Issue 3, pp. 255-265, 2015.
12. Xu, Rui, Donald Wunsch. "Survey of clustering algorithms", IEEE Transactions on Neural Networks, Vol. 16, Issue 3, pp. 645-678, 2012.