A Survey on Clustering Algorithms for Partitioning Method

Hoda Khanali
Department of Industrial Engineering,
Central Tehran Branch, Islamic Azad
University, Tehran, Iran

Babak Vaziri
Department of Industrial Engineering,
Central Tehran Branch, Islamic Azad
University, Tehran, Iran

ABSTRACT

Clustering is 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. 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. 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.

General Terms

Data mining

Keywords

Clustering methods; Partition algorithms; Fuzzy C-Means

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 the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters [2].

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].

The study is structured as follows: In Section 2, the clustering methods are reviewed, and partitioning clustering algorithms are described. Then in Section 3, these methods are generally compared. In the next section improved partitioning algorithms are analyzed, and then this analysis is assessed in Section 5. Finally the concluded points are presented in Section 6.

2. 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.

2.1 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.

2.2 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]. The algorithms in this category are DBSCAN [12], OPTICS [13], DENCLUE [14], CLIQUE [15], and so on.

2.3 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 multiresolution grid data structure. The main advantage of the approach is its fast processing time [2]. The algorithms in this category encompass STING [16], CLIQUE [15], and so on.

2.4 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. Typically, these algorithms are noniterative 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. 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]. The algorithms in this category are Cobweb [17], Classit [18], and so on.

2.5 Partitional clustering

Given the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another [2]. The general objective is to obtain the partition that, for a fixed number of clusters, minimizes the total square-error. Every partitional clustering algorithm obtains a single partition of the data instead of the clustering structure, such as a dendrogram. Thus, partitional methods have the advantage in applications involving large data sets for which the construction of a dendrogram is computationally very complex [8]. The commonly algorithms used this methods are k-means [19], k-medoids [20], Partitioning Around Medoids (PAM) [20], Clustering LARge Applications (CLARA) [21], Clustering Large Applications based on RANdomized Search (CLARANS) [22], Fuzzy C-Means (FCM) [23], and so on.

2.5.1 Partitional 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 [24]. 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 nonconvex 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 [25]. 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 [26]. 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.

3. COMPARISON OF CLUSTERING METHODS

According to the above, all algorithms are designed to minimize intracluster distances, and to maximize intercluster 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

Data size is one of the major issues of clustering such that incremental clustering is just applicable to large data sets according to Table 1. Meanwhile, some clustering methods can only identify a specific data shape such as spherical clusters in hierarchical and partitional clustering, complex and non-spherical clusters in density-based clustering, network of pattern space in grid-based clustering, and unlimited clusters in incremental clustering. Moreover, hierarchical clustering is based on distance, density and continuity, density-based clustering is based on a density of data points, grid-based clustering is based on network structure, incremental clustering is based on big data, and partitional clustering is based on distance. In the following Table 1, the advantages and disadvantages of each method are listed according to the above.

4. IMPROVED PARTITION ALGORITHMS

Multi-center Fuzzy C-means algorithm based on Transitive Closure and Spectral Clustering (MFCM-TCSC) [27] 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. On the other the problem of clustering samples is changed into a problem of merging subclusters, thus the computational load is low, and has strong robustness.

Robust clustering approach [28] 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. 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. 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 [29] improves the performance of FCM clustering algorithm by combining it with Relaxed constraints support vector machine (RSVM) [30] 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.

Candidate Groups Search [31] designs a new scheme, candidate group search (CGS), to get better chance escaping from local optimum in one way and using less time in the other. First, CGS uses some selection rules to identify the candidate group set for each center. Then, it screens through all the data set. If it belongs to the candidate group, the center has to be replaced and using k-harmonic means (KHM) [32] to obtain a new solution.

SC-FCM [33] is a hybrid algorithm based on the principles of FCM and one of the newest optimization algorithms, stem cells algorithm (SCA) [34]. In this algorithm, an initial population is selected from the members and then in subsequent iterations new members are added to the old population by a specific percentage fraction. The competence is not a proprietary of a single cell but is dedicated to a group of cells in reaching at an ideal clustering (formation of a single or some special organs inspired by the SCA algorithm's nature). The important part of SC-FCM method is the ability of automatic inspection of the optimum number of clusters in a large-scale dataset, high accuracy, and high speed of convergence.

Relative Entropy Fuzzy C-Means (REFCM) [35] 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. 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.

MinMax K-Means [36] 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. 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.

Interval Type-2 Credibilistic Clustering (IT2CC) [37] 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. Interval type-2 fuzzy sets use for modeling the uncertainty of fuzzifier which the credibility degrees are transformed to interval type-2 form to handle different sources of uncertainty. After acquiring the membership degrees and the other formulations, the upper and lower bounds are obtained by a specific strategy to consider the uncertainty of this parameter. This parameter controls the amount of fuzziness of the final partition.

Bias-correction FCM (BFCM), Bias-correction Gustafson and Kessel clustering (BGK), and Bias-correction Inter-Cluster Separation (BICS) [38] are according to bias-correction method. These methods integrating a bias-correction with an updating equation adjust the effects of initializations on fuzzy clustering algorithms, and also use an updating equation for a weighted parameter to ensure that the bias correction decreases when the number of iteration increases. This updating equation can stabilize algorithms when the number of iterations gradually increases.

Multi-central general type-2 fuzzy clustering [39] mainly focuses on uncertainty associated with the cluster centers. A set of points is considered as the center for each cluster. In addition, the membership values to the clusters, namely primary and secondary variables, are defined as general type-2 fuzzy sets. Primary variable indicates the degree of belonging to the central objects, and the secondary variables indicate the degree of belonging of the central objects to the center of the cluster. By increasing the uncertainty to select the center, more than two objects could be the cluster centers, but the approach is not any fuzzy mathematics in order to improve the uncertainty modeling.

5. 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. 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. F-TCLUST algorithm does not provide an evaluation of performance of classification curves. FCM-RSVM 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. 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. 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.

Table 2. Evaluation of improved partition algorithms

Algorithm	Discussed challenges	solutions	disadvantages
MFCM-TCSC	- Sensitive to non-spherical shape of clusters - Sensitive to initial prototypes	- fuzzy clustering based on transitive closure and spectral clustering	- Providing redundancy
F-TCLUST	- Sensitive to noise and outliers	- Discards a fixed fraction of data	- No evaluation of performance of the classification curves
FCM-RSVM	- Low membership values of data points in clusters	- Using relaxed constraints of support vector machine classifier	- Sensitivity to over-fitting
CGS	- Sensitivity to initial starting points - Convergence to the local optimum	- Deterministic selection rules and recursively solving KHM	- Not suitable for large instances
SC-FCM	- Sensitive problems to initialize - Convergence risks of large data sets	- Using multidimensional and self-renewal properties of stem cells	- Need to relatively large memory
REFCM	- Sensitive to noise and outliers	- Using fuzzy clustering features and relative entropy	---
MinMax K-Means	- Sensitive to initialize	- Minimizing the maximum intra-cluster variance instead of the sum of the intra-cluster variances	- Not using kernel-based clustering
IT2CC	- The uncertainty of fuzzy membership degrees - Little attention to separation of clusters in the objective function	- Using type-2 fuzzy clustering	- High computational complexity - High computational time - High iterations

BFCM, BGK, BICS	- Sensitive to initialize	- Using bias-correction approach	- High iterations
Multi-central general type-2 fuzzy clustering	- The uncertainty for selecting cluster centers - Convergence to the local optimum (No optimization number of clusters)	- Using type-2 fuzzy clustering - Determining the degree of belonging to the multi-central clusters	- High computational complexity - High computational time - High iterations - Not applying fuzzy mathematics

6. 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. 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 partitional clustering in data mining, and its being widely in recent years, clustering algorithms have become into a field of active and dynamic research. Therefore, improving the partition clustering algorithms such as K-means and FCM could be an interesting issue for future research.

7. REFERENCES

[1] Abonyi, J. and Feil, B. 2007 Cluster Analysis for data Mining and System Identification. Birkhäuser Verlag AG.

[2] Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P. 1998 Automatic subspace clustering of high dimensional data for data mining applications.

[3] DeRosa, M. 2004 Data Mining and Data Analysis for Counterterrorism.

[4] Dunn, J. 1974 A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters.

[5] Ester, M., Kriegel, H. and Sander, J. 1996 A density-based algorithm for discovering clusters in large spatial databases.

[6] Ester, M., Kriegel, H., Sander, J. and Xu, X. 1996 A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

[7] Fisher, D. 1987 Improving Inference Through Conceptual Clustering.

[8] Fritz, H. and Garcia-Escudero, L. 2013 Robust constrained fuzzy clustering.

[9] Gennari, J., Langley, P. and Fisher, D. 1989 Models of incremental concept formation.

[10] Gorunescu, F. 2011 Data Mining Concepts, Models and Techniques. Springer-Verlag Berlin Heidelberg.

[11] Han, J. and Kamber, M. 2006 Data Mining: Concepts and Techniques. Elsevier Inc.

[12] Hinneburg, A. and Keim, D. 1998 An efficient approach to clustering in large multimedia databases with noise.

[13] Hogo, M. 2010 Evaluation of e-learning systems based on fuzzy clustering models and statistical tools.

[14] Huang, T. and Hsu, W. 2013 Conjecturable knowledge discovery: A fuzzy clustering approach.

[15] Hung, C., Chiou, H. and Yang, W. 2013 Candidate groups search for K-harmonic means data clustering.

[16] Izakian, H. and Pedrycz, W. 2015 Fuzzy clustering of time series data using dynamic time warping distance.

[17] János Abonyi, B. F. 2007 Cluster Analysis for data Mining and System Identification.

[18] Kantardzic, M. 2003 Data Mining Concepts, Models, Methods, and Algorithms. The Institute of Electrical and Electronics Engineers, Inc.

[19] Karypis, G., Han, E. and Kumar, V. 1999 CHAMELEON: A hierarchical clustering algorithm using dynamic modelling.

[20] Kaufman, L. and Rousseeuw, P. 1987 Clustering by means of Medoids.

[21] Kaufman, L. and Rousseeuw, P. 1990 Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc.

[22] Lu, Y. and Ma, T. 2013 Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data.

[23] MacQueen, J. 1967 Some Methods for classification and Analysis of Multivariable Observations.

[24] Malek Mohamadi Golsefid, S. and Fazel Zarandi, M. 2016 Multi-central general type-2 fuzzy clustering approach for pattern recognitions.

[25] Mitra, S. and Acharya, T. 2003 Data mining multimedia, soft computing and bioinformatics. John Wiley & Sons, Inc.

[26] Ng, R. and Han, J. 1994 Efficient and effective clustering method for spatial data mining.

[27] Rostam Niakan Kalhori, M. and Fazel Zarandi, M. 2015 Interval type-2 credibilistic clustering for pattern recognition.

[28] Sabzekar, M. and Naghibzadeh, M. 2013 Fuzzy c-means improvement using relaxed constraints support vector machines.

[29] Sabzekar, M., Yazdi, H. Y. and Naghibzadeh, M. N. 2011 Relaxed constraints support vector machines for noisy data.

[30] Suganya, R. and Shanthi, R. 2012 Fuzzy C- Means Algorithm- A Review.

[31] Taherdangkoo, M. and Bagheri, M. 2013 A powerful hybrid clustering method based on modified stem cells and Fuzzy C-means algorithms.

[32] Taherdangkoo, M., Yazdi, M. and Bagheri, M. 2011 Stem cells optimization algorithm.

[33] Tzortzis, G. and Likas, A. 2014 The MinMax k-means clustering algorithm.

[34] Wang, W., Yang, J. and Muntz, R. 1997 STING: A statistical information grid approach to spatial data mining.

[35] Yang, M. and Tian, Y. 2015 Bias-correction fuzzy clustering algorithms.

[36] Zarinbal, M., Fazel Zarandi, M. and Turksen, I. 2014 Relative entropy fuzzy c-means clustering.

[37] Zeng, S., Tong, X. and Sang, N. 2013 Study on multi-center fuzzy C-means algorithm based on transitiveclosure and spectral clustering.

[38] Zhang, B., Hsu, M. H. and Dayal, U. 1999 K-harmonic means – a data clustering algorithm. Technical Report Hewlett–Packard Laboratories.

[39] Zhang, T., Ramakrishnan, R. and Livny, M. 1997 BIRCH: an efficient data clustering method for very large databases.