ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

A Detailed Study and Analysis of different Partitional Data Clustering Techniques

Ms. Aparna K[1,] Dr. Mydhili K Nair[2]

Associate Professor, Dept. of MCA, BMS Institute of Technology, Bangalore, Karnataka, India[1]

Associate Professor, Dept. of ISE, M S Ramaiah Institute of Technology, Bangalore, Karnataka, India[2]

Abstract: The concept of Data Clustering is considered to be very significant in various application areas like text mining, fraud detection, health care, image processing, bioinformatics etc. Due to its application in a variety of domains, various techniques are presented by many research domains in the literature. Data Clustering is one of the important tasks that make up Data Mining. Clustering can be classified into different types such as partitional, hierarchical, spectral, density-based, grid-based, model based etc. Among the different types of clustering available, partitional clustering is the most widely used in most of the applications since the computation involved is not very complex. Hence lot of research has been carried out in clustering using partitional method. In this paper, it is proposed to do a comprehensive study of the different partitional clustering techniques used in the literature which will also provide an insight into the recent problems in the same area. In this paper, sixteen research articles have been taken which are published by different publishers between the years 2005 and 2013. Various algorithms come under partitional clustering among which Bisecting K-Means is an excellent one that gives a good quality output for clustering large number of data. Also a broad analysis is carried out to provide an insight into the importance of the various approaches which can in turn throw light to developments in the same area.

Keywords: Data Mining, Clustering Techniques, Partitional clustering, Bisecting K-Means algorithm

I. INTRODUCTION

Clustering is an unsupervised learning method unlike the classification method which is generally viewed as a supervised learning technique. In other words, clustering groups the data based only on the information that is available in the dataset without any labels [1]. Clustering techniques can be generally classified into Partitional methods, Hierarchical approaches, Density-based algorithms, Probabilistic methods, Grid-based methods, Graph theory, Model-based approaches and so on [2]. Many partitional algorithms are recently introduced which are based on the technique of Evolutionary programming that includes Genetic Algorithms (GA), evolved from the Darwinian Theory.

Partitional clustering algorithms determine the clusters in such a way that the similarity within the clusters is maximum and the dissimilarity between the clusters is minimum. Though the K-Means algorithm is the most widely used algorithm under the partitional clustering because of its easy implementation factor, it has certain limitations. It does not give efficient results with differently shaped clusters [3] and moreover, it arbitrarily converges to local optima. But the clustering performance can be improved in terms of accuracy by incorporating constraints [4].

A. General Classification of Partitional Data Clustering

In this general architecture, the partitional data clustering techniques are classified into three main categories. They are partitional clustering, constraint based partitional clustering and evolutionary programming based clustering techniques. The partitional clustering is further subdivided into two categories which are K-Means method and other partitional clustering algorithm

Copyright to IJIRSET

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

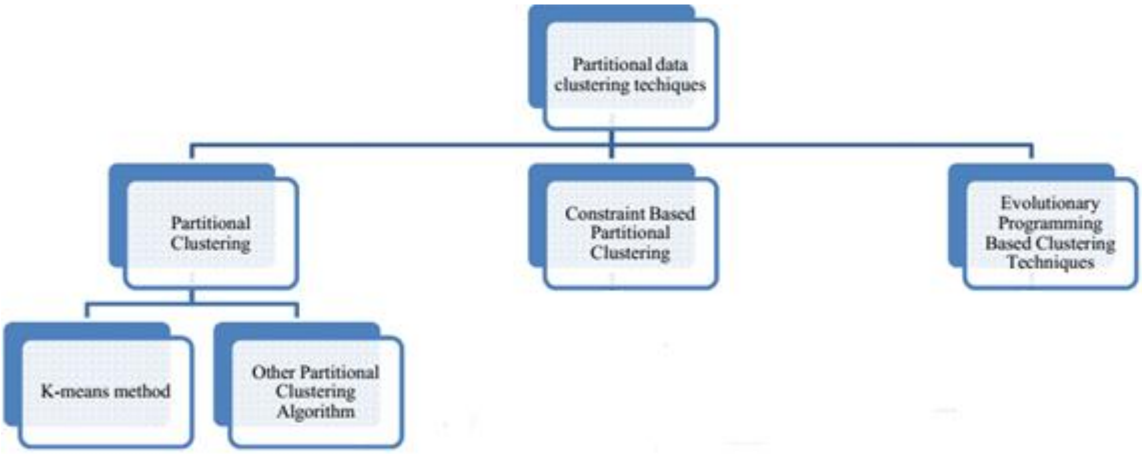

Fig. 1. Classification of Partitional Clustering Algorithms

II. LITERATURE SURVEY OF DIFFERENT PARTITIONAL DATA CLUSTERING TECHNIQUES

Partitional Clustering

Partitional clustering is further classified into K-Means method and based on other partitional clustering algorithms. Under this category sixteen research articles from the year 2005-2013 are taken and used for survey.

A. K-Means Method

Seven articles from the year 2005-2013 are taken and are used for survey for this category. In [5], Taher Niknam and Babak Amiriv have proposed a hybrid evolutionary algorithm called FAPSO-ACO-K (Fuzzy Adaptive Particle Swarm Optimization – Ant Colony Optimization – K-Means) to solve the nonlinear partitional clustering problem. The performance of this algorithm was evaluated through several benchmark datasets. The simulation results showed that the performance of this algorithm was better than the other traditional algorithms such as PSO (Particle Swarm Optimization), ACO, Simulated Annealing and so on.

In [6], M.Arshad designed a clustering algorithm based on KEA (Key phrase Extraction Algorithm) to solve the problem of document clustering in traditional clustering technique. The Kea Bisecting K-Means clustering algorithm was used to extract the test documents from a large amount of text documents in an easy and efficient way. The clustering algorithm was applied in order to generate the clustering document based on the extracted keys. The documents were grouped into several clusters like in the Bisecting K-Means algorithm. The results and the performance showed a consistently good quality of clusters which demonstrated in turn that the Bisecting K-Means is an excellent algorithm for clustering a large number of documents.

Based on the distance measure and K-Means algorithm, Dimitrios Charalampidis in [7], designed a Circular K-Means (CK-Means) algorithm. The objective of this algorithm was to design cluster vectors which contained directional information such as Fd, in a circular-shift invariant manner. In order to reduce the computational complexity, an efficient Fourier domain was represented in Circular K-Means algorithm. The information from the original feature vectors were not rejected in the designed algorithm when it performed clustering in a circular invariant manner. To estimate the correct number of clusters and to avoid local minima, a split and merge method (SMCK-Means) for CK-Means technique was designed. The designed algorithm was robust in terms of PCC (Percentage of Correct Clustering) and the estimated correct number of clusters.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

Imbalanced data always affected the performance of the representative algorithms which includes the hard K-Means and the fuzzy K-Means. Clusters of relatively uniform sizes were frequently produced even when input data have varied a cluster size; it was called as "uniform effect". The causes of this effect was analyzed by Jiye Liang *et al*, [8] and they projected that it possibly occurred more in the fuzzy K-Means clustering process than the hard K-Means clustering process. A multicenter clustering algorithm was designed in order to avoid this effect. The multicenters in this algorithm was used to represent each cluster rather than one single center. The effectiveness of the designed clustering algorithm in clustering balanced and imbalanced data was illustrated by the experimental results which were compared with the synthetic and real datasets.

An algorithm called K-MICA was proposed by Taher Niknam *et al*, [9] by combining MICA (Modify Imperialist Competitive Algorithm) and K-Means. This combined hybrid algorithm was used to optimize the clustering of N objects into K clusters. The performance was evaluated by testing it on several datasets and by comparing it with other clustering algorithms. The main focus was on the K-Means clustering algorithm. The initial choice of the cluster centers heavily affected the outputs of the K-Means algorithm. The hybrid evolutionary hybrid algorithm was tested on several datasets to overcome this drawback and its performance was compared with several existing traditional algorithms. The convergence of the hybrid algorithm to the global optimum solution was better than that of the other evolutionary algorithms.

In order to preprocess the document in document clustering, B S Vamsi Krishna et al [10], have applied the derived background knowledge from WordNet. Document vectors constructed from WordNet Synsets were used as input for clustering. A comparison was made between K-Means and Bisecting K-Means algorithms which showed that the Bisecting K-Means clustering algorithm was better than the standard K-Means clustering technique.

Three modified conventional moving K-Means clustering algorithms have been recommended by Nor Ashidi Mat Isa *et al*, [11] for the application of image segmentation. These three algorithms are fuzzy moving K-Means, adaptive moving K-Means and adaptive fuzzy moving K-Means algorithms. Standard images and hard evidence on microscopic digital image were used to analyze these algorithms. The segmentation result was compared with the conventional K-Means, fuzzy C-Means and moving K-Means algorithms. By qualitative and quantitative analysis, it was proved that this algorithm was less sensitive to noise and also the problems such as dead centers, center redundancy and trapped center at local minima were avoided. It was also illustrated that the above three modified algorithms were suitable to implement consumer electronics products based on their simplicity and capability.

B. Other Partitional Clustering Algorithms

Nine standard articles are chosen and surveyed under this category. In [12], Jian Yu has made two expectations about partitional clustering algorithm in order to study about the general C-Means algorithm and also the undesirable solution of GCM (General C-Means Clustering Model) is defined. The first assumption was that each subset be clustered into c ($c > 1$), which was expected to have a different prototype than others. The second assumption was that at a fixed point of GCM, the Undesirable Solutions of GCM (USGCM) must not be stable. The Hessian matrix of the objective function of GCM was obtained to find whether the USFCM (Undesirable Solutions of Fuzzy C-Means) was attractive at a stable point. The GCM model produced an iteration sequence which proved that it cannot be converged to a local minimum. This result was helpful to understand those partitional clustering algorithms which were included by the GCM model.

A temporal data clustering framework is presented by Yun Yang and Ke Chen [13]. A weighted clustering ensemble of multiple partitions which was produced by initial clustering analysis on different temporal data representation was used for this framework. A novel weighted consensus function was designed to reconcile the initial partitions to candidate consensus partitions from different perspectives. Also they introduced an agreement function to reconcile these candidate consensus partitions to a final partition. Therefore, the introduced weighted clustering ensemble algorithm provided an effective enabling technique for the joint use of different representations. It also reduced the information loss in a single representation and attained the various information sources which underlies in temporal data. Their

IJIRSET

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

method lead to capture the intrinsic structure of a data set. The method was evaluated with several benchmark datasets and showed that their method obtains best result for a variety of temporal data clustering tasks.

A clustering algorithm called LDA (Linear Discriminant Analysis) – based clustering (FLDC) algorithm based on Fisher criterion has been designed by Cheng-Hsuan Li *et al* [14]. The special cases of their designed unsupervised scatter matrices were the scatter matrices of LDA. The experimental results were compared both on synthetic and real datasets and the results showed that the designed algorithm outperformed the KMS (K-Means), KMD (K-Mediod), FCM (Fuzzy C-Means), GK (Gustafson-Kessel), GG (Gath Geva) and so on.

In [15], R. Indhumathi and S. Sathiyabama have designed a method which combines the dimensionality reduction through PCA (Principal Component Analysis) and Bisecting K-Means algorithm which the basic K-Means algorithm. Before clustering the original data, the PCA was applied on it. But the initialization of centroids affected the clustering results. A substantial improvement was seen by the experimental results in terms of processing time and accuracy of the clusters which reduced the dimension and initial centroid selection by PCA. The results also showed that the designed method outperformed K-Means.

The computational cost for large datasets was a major drawback of Fuzzy C-Means and it was unreliable for one index to determine the number of clusters, because the index was difficult to define. In order to overcome these problems and to identify the number of clusters simultaneously for a data set, a selection model was proposed in [16] by Yaonan Wang, Chunsheng Li, and Yi Zuo. This model combines multiple pairs of a fuzzy clustering algorithm and cluster validity index. The unreliability of traditional fuzzy clustering algorithm led to the design of this model. The method prevented FCM (Fuzzy C-Means) and its variants from local minima and drives them towards good clustering. Their method combined multiple pairs of algorithms and indexes in order to identify the number of clusters and to select the best clustering. The result of their selection model was relatively more reliable than that of a single pair of algorithm and index. The experiments on five real dataset revealed that FCM and AFCM (Alternative FCM) were relatively better and that of PFCM (Possibilistic FCM), which does not behave well.

E N Nasibov, G. Ulutagav in [17], have developed a Fuzzy Joint Points (FJP) method to deal with the fuzzy clustering problem in spatial data and the FJP method is based on the heuristic approach. The FJP algorithm transforms the distance-based approach of clustering to the fuzzy level sets-based approach. This makes is more possible to maintain detailed research by using fuzzy relation, fuzzy distance etc. The fundamental stages of a clustering process which includes determination of initial clusters, cluster validity and direct clustering are combined in FJP. This results in reduction in the total calculation time of a clustering process and hence can be used as a supplementary algorithm to cluster validity in FCM algorithm. This method also proved to be robust through noises.

The Fuzzy C-Means (FCM) algorithm is the most popular method among the fuzzy clustering methods used in image segmentation because it has robust characteristics for ambiguity and can retain much more information than hard segmentation methods. But the drawback of general Fuzzy C-Means algorithm is that it is very sensitive to noise and other imaging artefacts. Stelios Krinidis *et al* in [18] proposed a new clustering algorithm called FLICM (Fuzzy Local Information C-Means) in order to overcome the disadvantages of the Fuzzy C-Means algorithm and also to enhance the clustering performance. The major characteristic of FLICM was the use of a fuzzy local similarity measure which aimed to guarantee noise insensitiveness and image detail preservation. When the experiments were performed on synthetic and real-world images, the algorithm was very effective and efficient by remaining robust to noisy images. Also, this algorithm is completely independent of empirically adjusted parameters that are incorporated into all other Fuzzy C-Means algorithms in the literature.

The limitations of FCM were it requires the user to pre-define the number of clusters and different values of clusters correspond to different fuzzy partitions. In [19], Fuhua Yu *et al* have developed an improved automatic FCM clustering algorithm in order to overcome these disadvantages. The deviation coefficient and the distance coefficient were introduced into the developed automatic FCM algorithm which improved the robustness of the algorithm. The computational results showed that the improved automatic FCM algorithm performs excellently in the cluster validity.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

Generally clustering algorithm has extensive application in many fields. Due to the increased attention of clustering categorical data, and to minimize the within-cluster dispersion and to enhance the between-cluster separation, Liang Bai *et al*, [20] proposed a fuzzy clustering algorithm for categorical data which was an extension of the fuzzy K-Modes algorithm. They have rigorously derived the updating formulas of the membership matrix and the set of cluster prototypes in the clustering process, and proved the convergence of the algorithm under the optimization framework. The time complexity of the algorithm was analyzed, which was linear with respect to the number of data objects and attributes. Using several real data sets from UCI, the algorithm was tested and found that the algorithm was effective in clustering categorical data sets.

III. COMPARISON TABLE

This section presents a comparative analysis about partitional data clustering techniques methods available in the literature as shown in Table 1 below. Sixteen research articles are selected under the Partitional Clustering. Among this, most of the partitional clustering technique used here such as adaptive fuzzy moving K-Means and FLICM was less sensitive to noise. But these techniques required high processing time because of multiple evaluation performed to obtain better output.

The computational time of the clustering techniques such as K-MICA was very less because of the increased number of constraints. But the major drawback of this technique was its difficulty to cluster high dimensional data sets. However, even for high dimensional datasets the techniques such as KEA Bisecting K-Means determines the optimal number of clusters.

The clustering techniques such as PCA and Bisecting K-Means were robust because of the improved running time and accuracy. But these clustering techniques had a major drawback, which required pre-processing to improve its accuracy. The clustering methods such as FAPSO-ACO-K and K-MICA require prior information for efficient clustering. The Bisecting K-Means clustering algorithm used here provided better result than regular K-Means. However, most of the used ones here was time consuming because multiple evaluations were performed to obtain better output. Some of the clustering algorithms were not suitable for clustering high dimensional data sets.

Table 1: Comparison of all the papers considered in Literature

Types	Paper no.	Techniques	Advantages	Drawbacks
Partitional clustering	**19.** Fuhua Yu, Hongke Xu, Limin Wang, Xiaojian Zhou, (2010)	------	------------	-----------
	20. Liang Bai, JiyeLiang, ChuangyinDang and FuyuanCao, (2013)	------		
	5.Taher Niknam and Babak Amiri, (2010)	FAPSO-ACO–K	Less number of function evaluation was required	It was applicable only when the prior number of clusters were known
	6. M. Arshad (2012)	KEA Bisecting K-Means	Excellency in Clustering large number of documents without prior information	Extracted key phrase was required
	7. Dimitrios Charalampidis (2005)	SMCK-Means	Texture application such as texture segmentation and generation of key vectors required for building databases was an major advantage	Fluctuation regarding the correct number of clusters in the given data set
	8. Jiye Liang, Liang Bai, Chuangyin Dang and Fuyuan	Multicenter (MC) clustering (FGFKM,	Only two parameters were needed to easily	Only linearly separable clusters in the input space

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

	Cao (2012)	BMP, GMC)	setup	was obtained
	9. Taher Niknam, ElaheTaherianFard, NargesPourjafarian and AlirezaRousta, (2011)	K-MICA	Fast convergence which prevented it from falling into local optima	It was applicable only when the prior number of clusters was known
	10. B.S.Vamsi Krishna, P.Satheesh and Suneel Kumar R., (2012)	K-Means and Bisecting K-Means	Bisecting K-Means gives better results for larger data sets than regular K-Means	Time complexity
	11. Nor Ashidi Mat Isa, Amy A. Salamah and Umi Kalthum Ngah, (2009)	Adaptive Fuzzy Moving K-Means	Less sensitivity to noise	Required high processing time
	12. Jian Yu (2005)	General C-Means	Divergence of the GCM model in some cases provided better result	Iteration sequence produced doesn't guarantee the convergence of local minimum
	13. Yun Yang and Ke Chen, (2011)	Weighted clustering ensemble	Robust	High computational complexity
	14. Cheng-Hsuan Li, Bor-Chen Kuo and Chin-Teng Lin, (2011)	LDA-based clustering (FLDC)	Highest mean clustering accuracy	Time consuming
	15. R.Indhumathi and S.Sathiyabama, (2010)	PCA and a bisecting K-Means	Improvement in running Time and accuracy	Pre-processing required to improve efficiency
	16. Yaonan Wang, Chunsheng Li, and Yi Zuo , (2009)	Center initialization method based on MST	The computation time was significantly reduced	It was prototype based
	17. Efendi N. Nasibov and Gozde Ulutagay, (2007)	-------	---------	----------
	18. Stelios Krinidis and Vassilios Chatzis, (2010.)	Fuzzy local information C-Means (FLICM)	Relatively noise independent	Time complexity

IV. CONCLUSION

A detailed survey of various partitional techniques is presented. In this paper, sixteen articles are identified from 2005 to 2013 related to partitional data clustering techniques. Further classification is done based on different techniques and utilization and application of clustering. It is seen from the categorization, most of the techniques improved the robustness of the previously used clustering algorithms. Also, most of the clustering algorithms are time consuming because of multiple evaluations required for obtaining better output. It can also be inferred from here that clustering high dimensional data is not much effective in several techniques used here.

REFERENCES

[1] Shunzhi Zhu, Dingding Wang and Tao Li, "Data clustering with size constraints", Knowledge-Based Systems, Vol. 23, No. 8, pp. 883–889, 2010.

[2] Xueping Zhang, Jiayao Wang, Fang Wu, Zhongshan Fan and Xiaoqing Li, "A Novel Spatial Clustering with Obstacles Constraints Based on Genetic Algorithms and K-Medoids", Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications, Vol.1, pp. 605 – 610, 2006.

[3] Thiemo Krink and Sandra Paterlini, "Differential Evolution and Particle Swarm optimization in Partitional Clustering", Journal of Computational Statistics and Data Analysis, Vol. 50. No. 5, 2006.

[4] Ian Davidson, Kiri L. Wagstaff, and Sugato Basu, "Measuring Constraint-Set Utility for Partitional Clustering Algorithms", In: Proceedings of the Tenth European Conference on Principles and Practice of Knowledge Discovery in Databases, Vol. 4213, pp. 115-126, 2006.

[5] Taher Niknam and Babak Amiri, "An efficient hybrid approach based on PSO, ACO and K-Means for cluster analysis", Applied Soft Computing, Vol.10, No.1, pp.183-197, 2010.

[6] M. Arshad, "Implementation of Kea-Key phrase Extraction Algorithm by Using Bisecting K-Means Clustering Technique for Large and Dynamic Data Set", International Journal of Advanced Technology & Engineering Research, Vol.2, No.2, 2012.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

[7] Dimitrios Charalampidis, “A Modified K-Means Algorithm for Circular Invariant Clustering”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 12, December 2005

[8] Jiye Liang, Liang Bai, Chuangyin Dang and Fuyuan Cao, “The *K*-Means-Type Algorithms versus Imbalanced Data Distributions”, IEEE Transactions on Fuzzy Systems, Vol. 20, No. 4, 2012

[9] Taher Niknam, ElaheTaherianFard, NargesPourjafarian and AlirezaRousta,” An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering”, Engineering Applications of Artificial intelligence, Vol.24, No.2, pp.306-317, 2011

[10] B.S.Vamsi Krishna, P.Satheesh and Suneel Kumar R.,“ Comparative Study of K-means and Bisecting k-means Techniques in Word net Based Document Clustering”, International Journal of Engineering and Advanced Technology (IJEAT), Vol. 1, No.6, 2012.

[11] Nor Ashidi Mat Isa, Amy A. Salamah and Umi Kalthum Ngah, “Adaptive Fuzzy Moving K-means Clustering Algorithm for Image Segmentation”, IEEE Transactions On Consumer Electronics, Vol.55, No. 4, pp.2145-2153, 2009.

[12] Jian Yu, “General C-Means Clustering Model”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, August 2005

[13] Yun Yang and Ke Chen, “Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations”, IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 2, February 2011

[14] Cheng-Hsuan Li, Bor-Chen Kuo and Chin-Teng Lin, “LDA-Based Clustering Algorithm and Its Application to an Unsupervised Feature Extraction”, IEEE Transactions on Fuzzy systems, Vol. 19, No. 1, 2011

[15] R.Indhumathi and S.Sathiyabama, “Reducing and Clustering High Dimensional Data through Principal Component Analysis”, International Journal of Computer Applications, Vol.11, No.8, pp.0975-8887, 2010.

[16] Yaonan Wang, Chunsheng Li, and Yi Zuo , “A Selection Model for Optimal Fuzzy Clustering Algorithm and Number of Clusters Based on Competitive Comprehensive Fuzzy Evaluation”, IEEE Transactions On Fuzzy Systems, Vol. 17, No. 3, June 2009.

[17] Efendi N. Nasibov and Gozde Ulutagay, “A new unsupervised approach for fuzzy clustering”, Fuzzy Sets and Systems, Vol.158, No.19, pp. 2118– 2133, 2007

[18] Stelios Krinidis and Vassilios Chatzis, "A Robust Fuzzy Local Information C-means Clustering Algorithm", IEEE Transactions on Image Processing, vol. 19, No. 5, pp: 1328-1337, 2010.

[19] Fuhua Yu, Hongke Xu, Limin Wang, Xiaojian Zhou, “An Improved Automatic FCM Clustering Algorithm”, 2nd International Workshop on Database Technology and Applications (DBTA), pp. 1-4, 2010. Nov. 2010.

[20] Liang Bai, JiyeLiang, ChuangyinDang and FuyuanCao, “A novel fuzzy clustering algorithm with between-cluster information for categorical data”, Fuzzy Sets and Systems, Vol.215, pp.55-73, 2013.