Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 9, September -2017

Comparative Study of Initial Centroid based K-Mode Algorithm

Manisha Goyal [1], Shruti Aggarwal [2]

[1] Research Scholar, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib

[2] Assistant Professor, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib

Abstract — *Clustering is one of the techniques of the data mining, which defines classes and put the objects into one group having similar properties and objects having dissimilar properties into another group. An extension of the K-Means Algorithm, K-Mode Algorithm, is partitioning based clustering algorithm but it does not guarantee for the optimal solution. In this paper, there is the comparative analysis of Ini_Distance and Ini_Entropy Algorithm with Cao's methods, WK-Mode with Chan's Algorithm, Harmonic K-Mode with K-Mode and EC K-Mode Algorithm on real datasets. These algorithms are based on the selection of initial centroids in which the clustering accuracy is improved. The algorithms discussed in this study can be improved further by other better optimization techniques through research made in this field.*

Keywords- *Data Mining, Clustering, K-Means Algorithm, K-Mode Algorithm*

I. INTRODUCTION

Data mining is a process of extraction of the useful information and patterns from huge data. The knowledge to be mined have, within millions of objects is described by tens, hundreds or even thousands of different types of giving attributes or variables [1]. Due to the importance of extracting the information from the large data repositories, data mining has become an essential component, where the data could be stored in databases, data warehouses, or other information repositories [2]. The main purpose of this technique is to find patterns that were previously unknown and once these patterns are discovered, they can further be used to make certain decisions. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or pattern analysis. It consists of extract, transform, and load transaction data onto the data warehouse system, store and manage the data in a multidimensional database system, provide data access to business analysts and information technology professionals, analyse the data by application software, presents the data in a useful format, such as a graph or table [3]. The various application areas of data mining are marketing, education, banking, insurance, transportation, healthcare, finance etc.

Clustering is a technique of the data mining and is a form of unsupervised learning. It is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in another cluster [3]. Highly superior clusters have high intra-class similarity and low inter-class similarity. Several algorithms have been designed to perform clustering, each one uses different principles. The clustering algorithms can be generally classified into five categories that are described as follow:

1.1 Hierarchical Based Clustering: It builds a cluster hierarchy and is based on the connectivity approach. It uses the distance matrix criteria for clustering the data and constructs clusters step by step. The algorithms used in this approach are BIRCH, CURE etc.

1.2 Partitioning Based Clustering: It is a centroid based clustering in which data points splits into k partition and each partition represents a cluster. The amount of clusters for this technique should be predefined and the algorithms used in this approach are K-Means Algorithm, K-Medoid Algorithm, K-Nearest Neighbour Algorithm etc [4].

1.3 Density Based Clustering: This method finds the cluster according to the regions which grow with high density. The algorithms used in this approach are DBSCAN, GDBSCAN, OPTICS etc.

1.4 Grid Based Clustering: This method maps all the objects in a cluster into a number of square cells, known as grids. It has a fast processing time that depends on the size of the grid instead of the data. The algorithms used in this approach are STING, CLIQUE etc.

1.5 Model Based Clustering: In this method, each of the clusters is best fitted to the given model. It may locate clusters by constructing a density function that reflects the space distribution of the data points. The algorithm used in this approach is Expectation-Maximization (EM) algorithm [3].

K-Means Algorithm is a popular unsupervised learning algorithm and is a partitioning based algorithm for clustering. K data elements are selected as initial centers; then distance of all data elements is deliberate by Euclidean distance formula. Data elements having less distance to centroids are stimulated to the appropriate cluster. The process is undergone until no more alteration occurs in clusters [2].

An extension of K-Means Algorithm, K-Mode Algorithm, is the partitioning based clustering algorithm. The k-modes approach modifies the standard k-means process of clustering by replacing the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent cluster centres and updating modes in each of iterations of the clustering process [5].

Algorithm for K-Mode Clustering:

The steps of the K-Mode Algorithm are as follow:

INPUT: Number of desired clusters K, Data objects D= {d1, d2…dn}

OUTPUT: A set of K clusters

1. *Generate K clusters by randomly selecting the data objects and choose K initial cluster center, one for every of the cluster.*
2. *Assign a data object to the cluster whose cluster center is near toward it.*
3. *Update the K cluster base on allocation of data objects and calculate K latest modes of every one clusters.*
4. *Repeat step 2 to 3 awaiting no data object has changed cluster relationship otherwise some additional predefined criterion is fulfilled.*

The k-mode clustering algorithm is efficient in clustering the large categorical data, whereas k-means algorithm does not work well for categorical data and also produces locally minimal clustering results. It is famous for its simplicity and the efficiency of the clustering process is maintained.

II. LITERATURE SURVEY

K-Mode algorithm was proposed by Haung in 1997 in which the mode is calculated instead of calculating the mean value in order to find the accurate clusters to avoid overlapping to deal with categorical attributes.

One of the K-Mode Algorithm based upon cluster centres [6] in which a suitable value of regularization parameter was chosen to find the most stable clustering results and to control the number of clusters in the clustering process.

Another K-Mode Algorithm is Iterative K-Mode [7] introduced an initialization method based on Bradley's iterative initial-point refinement algorithm to the K-Modes clustering and results in accurate number of clusters.

Another one of the K-Mode based upon distance metrics [8] introduced a dissimilarity measure based on the distance between two attribute values of the same attribute and the similarity of two attribute values is dependent on their relationship with other attributes.

The other k-mode algorithm is COOLCAT algorithm [9] which is able to deal with clustering of data streams and is based on the notion of entropy. It depends on an input parameter that represents the size of the small cluster.

One more k-mode algorithm is Distance based k-mode (Cao's Method) [10] proposed an initialization method for categorical data and the distance between objects was calculated based on the frequency of attribute values. It calculates the densities of all the objects for categorical data and the process was limited to sub-sample datasets.

One more K-Mode Algorithm is based on Cost function [11] in which cost function added weight for numeric attributes computed from the dataset and all numeric attributes were normalized and discretised to do the calculations.

Another K-Mode Algorithm is DILCA Algorithm [12] proposed a method called Distance learning for Categorical Attributes. The distance between two values of a categorical attribute was determined by the way in which the value of the other attributes was distributed in the dataset.

DISC algorithm [13] is another one K-Mode Algorithm that suggested the method Data-Intensive Similarity Measure for Categorical Data. This measure didn't require any domain knowledge to understand the dataset.

Biological and Genetic taxonomy information based k-mode [14] is the other K-Mode Algorithm that proposed a new dissimilarity measure based on the idea of biological and genetic taxonomy and rough membership function and improved the accuracy of the clusters.

K-Mode based upon unified similarity metrics algorithm [15] proposed a penalized competitive learning algorithm and these algorithm required some initial value of number of clusters which should be greater than the original value of number of clusters. The resulting clusters are more accurate than the original K-Mode Algorithm.

Cluster centre initialization based k-mode algorithm [16] is one of the K-Mode Algorithm that introduced some objects which are very similar to each other and have same cluster membership irrespective of the choice of initial cluster centres and generate accurate clusters using prominent attributes.

Entropy based similarity coefficient k-mode algorithm (EC K-Mode) [17] is another K-Mode Algorithm that improved the cluster accuracy and analysed the time complexity while retaining the scalability of the K-Mode Algorithm.

Dissimilarity based k-mode [18] is also the one of the K-Mode Algorithm in which a new dissimilarity measure is proposed in which the modes of clusters were updated in each iteration ad utilizes some theorems to update the mode of the cluster.

Ini_Distance and Ini_Entropy Algorithm [19] presented two different initialization algorithms in k-mode where the first algorithm is based on the traditional distance-based outlier detection technique, and the second algorithm is based on the partition entropy-based outlier detection technique and guarantees that chosen initial cluster centers are not outliers. A new distance metric i.e. weighted matching distance metric is adopted to calculate the distance between two objects described by categorical attributes.

WK-Mode Algorithm [20] is the weighting k-mode algorithm used to automatically compute the weight of all the dimensions in each cluster by using complement entropy and used to identify the subsets of important dimensions that categories the different dimensions.

Chan's Algorithm [21] is an attribute weighting algorithm that generates the weight for each attribute from each cluster, which allows the use of the k-means type paradigm to efficiently cluster large data sets.
the tables.

III. ANALYSIS AND RESULTS

3.1 Datasets

The datasets that are mushroom, soyabean and vote from UCI Machine Learning Repository was used to analyse the performance. The detailed description of the datasets is discussed in the Table 1.

Properties	Soyabean	Mushroom	Vote
Number of Classes	4	2	2
Number of Instances	47	8124	435
Number of Attributes	35	24	16

Table 1: Properties of three UCI Datasets

Soyabean dataset consists of 4 numbers of classes, 47 numbers of instances and 35 numbers of attributes. Mushroom dataset consists of 2 numbers of classes, 8124 numbers of instances and 24 numbers of attributes. Vote dataset consists of 2 numbers of classes, 435 numbers of instances and 16 numbers of attributes.

3.2 Output Parameters

The result is compared using three output parameters that are accuracy, precision and recall.

- **True Positive Rate (TP):** A true positive test result is one that detects the condition when the condition is present.
- **True Negative Rate (TN):** A true negative test result is one that does not detect the condition when the condition is absent.
- **False Positive Rate (FP):** A false positive test result is one that detects the condition when the condition is absent.
- **False Negative Rate (FN):** A false negative test result is one that does not detect the condition when the condition is present.

The parameters based upon which the performance of the algorithms is evaluated and compared is described below:

i. Accuracy: The Accuracy is the total number of module that is predicted correctly.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad \text{…….. Eq (1)}$$

ii. Precision: Precision is the measure of exactness i.e. what percentage of tuples labeled as positive that are actually such.

$$\text{Precision} = \frac{(TP)}{(TP+FP)} \quad \text{…….. Eq (2)}$$

iii. Recall: Recall is the measure of completeness i.e. what percentage of positive tuples did the classifier label as positive.

$$\text{Recall} = \frac{(TP)}{(TP+FN)} \quad \text{..…... Eq (3)}$$

3.3 Comparison of Ini_Distance and Ini_Entropy Algorithm with Cao's methods [19]

Ini_Distance and Ini_Entropy Algorithm presented two different initialization algorithms where the first algorithm is based on the traditional distance-based outlier detection technique, and the second algorithm is based on the partition

entropy-based outlier detection technique that measures the degree of outlierness of each objects and better or equals results over cao's method.

3.3.1 Accuracy

The experimental results of the Cao's Method, Ini_Distance and Ini_Entropy Algorithm are shown in the Table 2.

Datasets	Cao's Method	Ini_Distance Algorithm	Ini_Entropy Algorithm
Soyabean	1	1	1
Mushroom	0.8754	0.8941	0.8876
Vote	0.8621	0.8690	0.8690

Table 2: Accuracy of Cao's Method, Ini_Distance and Ini_Entropy Algorithm

Ini_Distance and Ini_Entropy Algorithm show the better accuracy results than cao's method on mushroom and vote datasets and also show equal result on the soyabean dataset.

3.3.2 Precision

The experimental results of the Cao's Method, Ini_Distance and Ini_Entropy Algorithm are shown in the Table 3.

Datasets	Cao's Method	Ini_Distance Algorithm	Ini_Entropy Algorithm
Soyabean	1	1	1
Mushroom	0.9019	0.9138	0.9095
Vote	0.8571	0.8630	0.8630

Table 3: Precision of Cao's Method, Ini_Distance and Ini_Entropy Algorithm

Ini_Distance and Ini_Entropy Algorithm show the better precision results than cao's method on mushroom and vote datasets and also show equal result on the soyabean dataset.

3.3.3 Recall

The experimental results of the Cao's Method, Ini_Distance and Ini_Entropy Algorithm are shown in the Table 4.

Datasets	Cao's Method	Ini_Distance Algorithm	Ini_Entropy Algorithm
Soyabean	1	1	1
Mushroom	0.8709	0.8903	0.8835
Vote	0.8755	0.8811	0.8811

Table 4: Recall of Cao's Method, Ini_Distance and Ini_Entropy Algorithm

Ini_Distance and Ini_Entropy Algorithm show the better recall results than cao's method on mushroom and vote datasets and also show equal result on the soyabean dataset.

3.4 Comparison of WK-Mode with Chan's Algorithm [20]

WK-Mode Algorithm is presented for subspace clustering that automatically computed the weight of all dimension to each cluster by using complemented entropy. This algorithm shows better results than chan's algorithm.

3.4.1 Accuracy

Figure 1 shows that the experimental result of the wk-mode algorithm and chan's algorithm. WK-Mode Algorithm shows better results than the chan's algorithm over soyabean, mushroom and vote datasets.

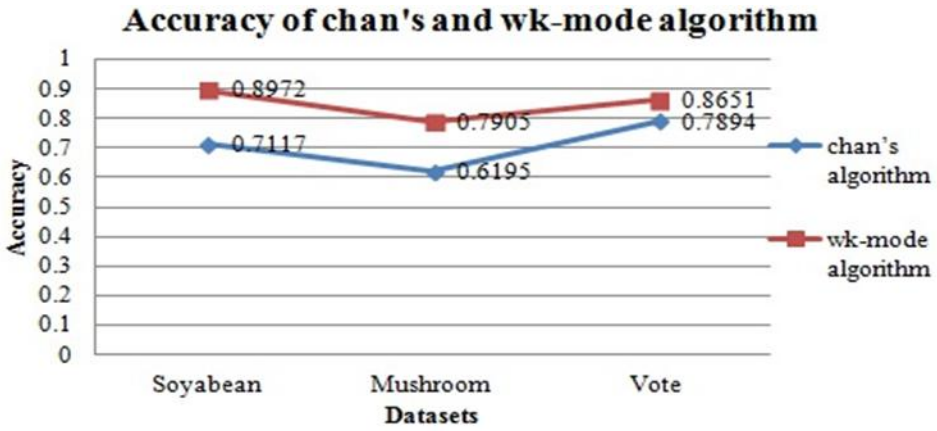

Fig 1: Accuracy of chan and wk-mode algorithm

3.5 Comparison of Harmonic K-Mode with K-Mode and EC K-Mode Algorithm

The harmonic k-mode algorithm is hybrid to find the optimal value of the centroids so that accurate clusters may be formed and shows better result than the K-Mode Algorithm and EC K-Mode Algorithm when number of cluster is 8.

3.5.1 Accuracy

The experimental result of the K-Mode, EC K-Mode Algorithm and Harmonic K-Mode Algorithm is shown in the Figure 2.

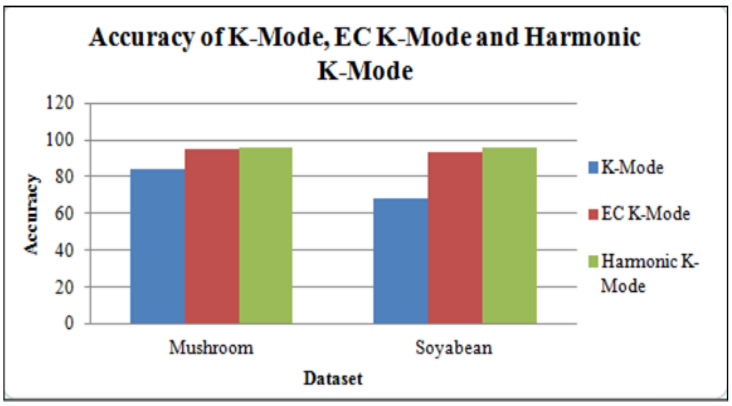

Fig 2: Accuracy of K-Mode, EC K-Mode and Harmonic K-Mode

Harmonic K-Mode Algorithm shows the better accuracy results than K-Mode and EC K-Mode Algorithm on mushroom and soyabean dataset.

3.5.2 Precision

The experimental result of the K-Mode, EC K-Mode Algorithm and Harmonic K-Mode Algorithm is shown in the Figure 3.

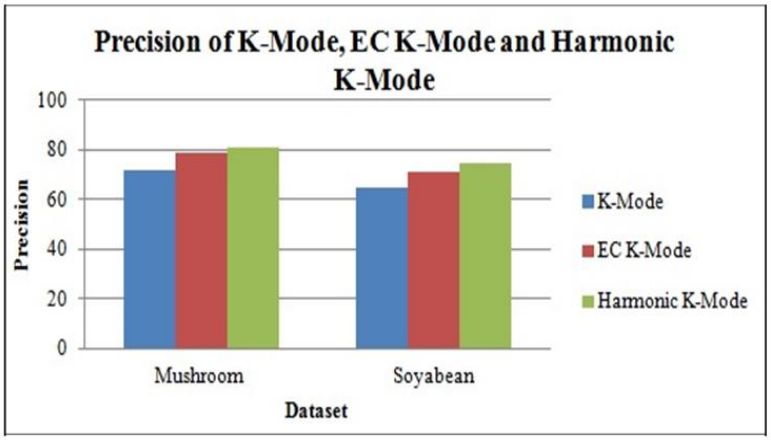

Fig 3: Precision of K-Mode, EC K-Mode and Harmonic K-Mode

Harmonic K-Mode Algorithm shows the better precision results than K-Mode and EC K-Mode Algorithm on mushroom and soyabean datasets.

3.5.3 Recall

The experimental result of the K-Mode, EC K-Mode Algorithm and Harmonic K-Mode Algorithm is shown in the Figure 4.

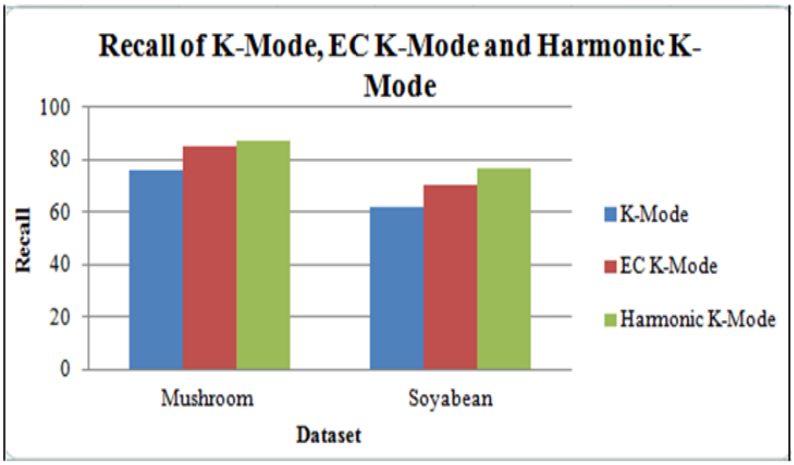

Fig 4: Recall of K-Mode, EC K-Mode and Harmonic K-Mode

Harmonic K-Mode Algorithm shows the better recall results than K-Mode and EC K-Mode Algorithm on mushroom and soyabean datasets.

The comparisons of the above algorithms are based on the selection of the initial centroids that improves the accuracy of the algorithm. The results show that Ini_Distance and Ini_Entropy Algorithm has better accuracy than Cao's Method, WK-Mode Algorithm has better accuracy than Chan's Algorithm, the Harmonic K-Mode Algorithm has better accuracy than K-Mode and EC K-Mode Algorithm.

IV. CONCLUSION

Clustering is one of the most extensively used data mining algorithm and an important topic of research. An extension of K-Means Algorithm, K-Mode Algorithm, is a popular clustering algorithm that uses simple matching dissimilarity function instead of Euclidean distance. In this paper, there is the comparative analysis of Ini_Distance and Ini_Entropy Algorithm with Cao's methods, WK-Mode with Chan's Algorithm, Harmonic K-Mode with K-Mode and EC K-Mode Algorithm on real datasets. These algorithms are based on the selection of initial centroids in which the clustering accuracy is improved. The results show that Ini_Distance and Ini_Entropy Algorithm has better accuracy than Cao's Method, WK-Mode Algorithm has better accuracy than Chan's Algorithm, the Harmonic K-Mode Algorithm has better accuracy than K-Mode and EC K-Mode Algorithm. The k-mode algorithm is still at the stage of exploration and the work can be carried out on the improvement of efficiency, to enhance cluster quality and to achieve more accuracy.

REFERENCES

[1]. Vinaya Sawant, Ketan Shah, "Performance Evaluation of Distributed Association Rule Mining Algorithms", 7th International Conference on Communication, Computing and Virtualization, Elsevier, Vol. 79, pp. 127-134, 2016.
[2]. Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector", 3rd International Conference on Recent Trends in Computing, Elsevier, Vol. 57, pp. 500-508, 2015.
[3]. Han, J. and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 3rd Edition, 2011.
[4]. Jeyhun Karimov, Murat Ozbayoglu, "Clustering Quality Improvement of k-means using a Hybrid Evolutionary Model", Conference Organized by Missouri University of Science and Technology, Elsevier, Vol. 61, pp. 38-45, 2015.
[5]. Md Anisur Rahman, Md Zahidul Islam, Terry Bossomaier, "ModEx and Seed-Detective: Two novel techniques for high quality clustering by using good initial seeds in K-Means", Journal of King Saud University – Computer and Information Sciences, Elsevier, Vol. 27, pp. 113-128, 2015.
[6]. O. M. San, V. Hyunh, Y. Nakamori, "An Alternative Extension of the k-Means Algorithm for Clustering Categorical Data". International Journal Applied Math and Computer Science, Vol.14, pp. 241–247, 2004.
[7]. Y. Sun, Q. Zhu, Z. Chen, "An iterative initial-points refinement algorithm for categorical data clustering", Pattern Recognition Letters, Elsevier, Vol. 23, Issue. 7, pp. 875–884, 2002.

[8]. Amir Ahmad, Lipika Dey, "A K-Mean Clustering Algorithm for Mixed Numeric and Categorical Data", Data & Knowledge Engineering, Science Direct, Vol. 63, pp. 503–527, 2007.
[9]. D. Barbara, J. Coute, Yi Li, "COOLCAT: An entropy based algorithm for categorical clustering", Proceedings of the eleventh international conference on Information and knowledge management, USA, ACM, pp. 582-589, 2002.
[10]. F. Cao, J. Liang, L. Bai, "A new initialization method for categorical data clustering", Expert Systems and Applications, Science Direct, Vol. 36, pp. 10223-10228, 2009.
[11]. Amir Ahmad, Lipika Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set", Pattern Recognition Letters, Science Direct, Vol. 28, Issue. 1, pp. 110–118, 2007.
[12]. D. Ienco, R. G. Pensa, R. Meo, "From Context to Distance: Learning Dissimilarity for Categorical Data Clustering", ACM Transactions on Knowledge Discovery from Data, pp.1-22, 2011.
[13]. A. Desai, H. Singh, V. Pudi, "DISC: Data Intensive Similarity Measure for Categorical Data", Proceedings of Advances in Knowledge Discovery and Data Mining – 15th Pacific Asia Conference, Springer, pp. 469 – 481, 2011.
[14]. F. Cao, J. Liang, D. Li, L. Bai, C. Dang, "A dissimilarity measure for the k-modes clustering algorithm", Knowledge-Based Systems, Elsevier, Vol. 26, pp. 120–127, 2012.
[15]. Y. M. Cheung, H. Jia, "Categorical and numerical attribute data clustering based on a unified similarity metric without knowing cluster number", Pattern Recognition, Elsevier, Vol. 46, pp. 2228–2238, 2013.
[16]. S. S. Khan, A. Ahmad, "Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering", Expert Systems with Applications, Elsevier, Vol. 40, pp. 7444–7456, 2013.
[17]. Ravi Sankar Sangam, Hari Om, "The k-modes algorithm with entropy based similarity coefficient", 2nd International Symposium on Big Data and Cloud Computing, Procedia Computer Science, Elsevier, Vol. 50, pp. 93-98, 2015.
[18]. Michael K. Ng, Mark Junjie Li, Joshua Zhexue Huang, "On the Impact of Dissimilarity Measure in K-Modes Clustering Algorithm", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, Issue. 3, pp. 503-507, 2007.
[19]. Feng Jiang, Guozhu Liu, Junwei Du, Yuefei Sui, "Initialization of K-modes clustering using outlier detection techniques", Information Sciences, Science Direct, Vol. 332, pp. 167-183, 2016.
[20]. F. Cao, J. Laing, D. Li, X. Zao, "A Weighting K-Modes Algorithm for subspace clustering of categorical data", Neurocomputing, Elsevier, Vol. 108, pp. 23-30, 2013.
[21]. E. Y. Chan, W. K. Ching, M. K. Ng, J. Z. Haung, "An Optimization Algorithm for clustering using weighted dissimilarity measure", Pattern Recognition, Elsevier, Vol. 37, Issue. 5, pp. 943-952, 2004.

