PCFA: Mining of Projected Clusters in High Dimensional Data Using Modified FCM Algorithm

Ilango Murugappan[1] and Mohan Vasudev[2]
[1]Department of Computer Applications, KLN College of Engineering, India
[2]Department of Mathematics, Thiagarajar College of Engineering, India

Abstract: *Data deals with the specific problem of partitioning a group of objects into a fixed number of subsets, so that the similarity of the objects in each subset is increased and the similarity across subsets is reduced. Several algorithms have been proposed in the literature for clustering, where k-means clustering and Fuzzy C-Means (FCM) clustering are the two popular algorithms for partitioning the numerical data into groups. But, due to the drawbacks of both categories of algorithms, recent researches have paid more attention on modifying the clustering algorithms. In this paper, we have made an extensive analysis on modifying the FCM clustering algorithm to overcome the difficulties possessed by the k-means and FCM algorithms over high dimensional data. According to, we have proposed an algorithm, called Projected Clustering based on FCM Algorithm (PCFA). Here, we have utilized the standard FCM clustering algorithm for sub-clustering high dimensional data into reference centroids. The matrix containing the reference values is then fed as an input to the modified FCM algorithm. Finally, experimentation is carried out on the very large dimensional datasets obtained from the benchmarks data repositories and the performance of the PCFA algorithm is evaluated with the help of clustering accuracy, memory usage and the computation time. The evaluation results showed that, the PCFA algorithm shows approximately 20% improvement in the execution time and 50% improvement in memory usage over the PCKA algorithm.*

Keywords: *Clustering, FCM, Modified FCM, k-mean clustering, accuracy, memory usage, computation time.*

Received November 29, 2011; accepted May 22, 2012; published online January 29, 2013

1. Introduction

Clustering is one prominent method in data mining field, the main reason for its reputation is the capability to work on datasets with smallest amount or without apriori facts [2]. In recent times, high-dimensional data has stimulated the attention of database researchers owing to its new challenges fetched to the society. The distance from a record to its adjacent neighbor can come up to its distance to the farthest record in high dimensional space [12]. In high-dimensional data, clusters can exist in subspaces that put out of sight themselves from conventional clustering techniques [22]. Because of the dilemma of dimensionality, most clustering algorithms will not perform proficiently in high-dimensional spaces [1, 8]. In a high-dimensional space, the distance between each pair of points is approximately identical for a wide range of data distributions and distance functions [4]. Feature selection procedures are generally made use of as a preprocessing stage for clustering, so as to triumph over the problem of dimensionality. The most relevant dimensions are chosen by discarding unrelated and unnecessary ones. Sush methods accelerate clustering algorithms and develop their performance [5].

The clustering accuracy can be gravely tarnished if inaccurate values are taken [17]. In real circumstances, it is hardly ever possible for users to provide the parameter values correctly, which brings about practical complexities in relating these algorithms to real data [13]. Merely, a subset of attributes is appropriate to each cluster, and those clusters can be able to have a diverse set of significant attributes [18]. An attribute is appropriate to a cluster if it aid to spot the member records of it [3]. Searching clusters and their appropriate attributes from a dataset is called as the projected (subspace) clustering. A projected clustering algorithm establishes a set of attributes that it supposes to be most significant for each cluster. We name those attributes the “selected attributes" of the cluster [23].

The distance of any two points in a high dimensional space is approximately identical for a large class of frequent distributions. Conversely the extensively employed distance measures are highly significant in subsets or so called projections of the high dimensional space, where the object values are dense [19]. Projected clusters can be formed in a variety of data. Projected clustering has been triumphant in a computer vision task and has impending applications in e-commerce [25]. The projections of its members will be concentrated on a small range of values that encloses few or no projections of other objects, while all objects are projected onto an appropriate dimension of a cluster [26].

Jhimli *et al.* [16] proposed an algorithm for clustering items in different data sources. They have proposed the idea of best cluster by bearing in mind that average degree of variation of a class. Also, they have planned an alternative algorithm to find best cluster among items in different data sources. Gabriela *et al.* [14], evaluated systematically state-of-the-art subspace and projected clustering techniques under a wide range of experimental settings. They have discussed the observed performance of the compared techniques, and they made recommendations regarding what type of techniques were suitable for what kind of problems.

The above mentioned researches have given the idea of proposing a method, which provide a solution for clustering in high dimensional data. We proposed a new approach that will help in the clustering of high dimensional data by reducing the dimensions without losing the data. The proposed approach is a combination of standard Fuzzy C-Means (FCM) algorithm and a modified FCM. The initial process in our proposed approach is managed by the standard FCM algorithm; the process is a gridding process to concentrate similar data into blocks. Gridding generated some centroids, which represent the blocks of data. The next processes are relevant attribute analysis and outlier detection, which are used for the removal of redundant data from the data set which are provided. The above process helps to increase the speed of execution of our proposed method by reducing the space complexity. A binary data corresponding to the input data is generated as a result of the relevant attribute analysis. The modified FCM uses the binary data for the processing of the membership functions, which are used for the updation of new centroids. There have been many researches related to the projected cluster mining such as [7, 9, 15, 21, 22, 24,].

The main contribution of the paper is discussed as follows:

- Design an algorithm called, Projected Clustering based on FCM Algorithm PCFA, a modified FCM approach for mining of projected clusters from high dimensional data.
- Design a methodology to carry the attribute relevant analysis that is used to detect the outlier and finding projected clusters.
- Adapting standard FCM algorithm to grid the data points for dimensionality reduction.

The rest of the paper is organized as follows: The motivating algorithms are given in section 2. The proposed PCFA algorithm is presented in section 3. The experimental results and analysis of the proposed approach are given in section 4. Conclusion is summed up in section 5.

2. Motivating Algorithms

This section explains the discussion of the methods, which motivated us to develop a new method for high dimensional data clustering. The main method behind our proposed approach is the FCM algorithm [6, 11], which is used for the clustering of data. In our proposed approach, we modify the standard FCM by adding binary values. The key idea of our proposed approach is obtained by studying the algorithm, PCKA, by Mohamed Bouguessa and Shergrui Wang [20]. In their approach, they proposed an approach to mine the projected clusters from the high-dimensional data. The projected cluster is a subset of data in the dataset and has a particular dimension.

- *PCKA Algorithm*

The PCKA algorithm [20], is an algorithm to find the projected clusters from the dataset provided for clustering. Let S be a dataset of d-dimensional points, where, the set of attributes is denoted by $A=\{A_1,A_2,...A_d\}$. Let $X=\{x_1,x_2,...x_N\}$ be the set of N data points, where $x_{ij}=(x_{ij},...x_{ij},...x_{id})$. Each $x_{ij}(i=1,... N; j=1,... d)$ corresponds to the value of data point x_i on attribute A_j . x_{ij} is called as a 1-d point. Assume that each data point x_i belongs either to one projected cluster or to the set of outliers Z. Given the number of clusters n, which is an input parameter, a projected cluster $C_y,y=1,...n$ is defined as a pair (P_y, D_y), where P_y a subset of data is points of S and D_y is a subset of dimensions of A, such that the projections of the points in P_y along each dimension in D_y are closely clustered. The dimensions in P_y are called relevant dimensions for the cluster C_y. The remaining dimensions i.e., $A\text{-}D_y$, are called irrelevant dimensions for the cluster C_y. The cardinality of the set D_y is denoted by d_y, where $d_y \leq d$ and n denotes the cardinality of the set P_y, where $n_y \leq N$. PCKA assist in finding out axis-parallel projected clusters which fulfills the following properties:

1. Projected clusters should be dense. Particularly, the projected values of the data points along each dimension of $\{Dy\}_{y=1,...n}$ form regions of high density in contrast to those in every dimension of $\{A\text{-}D_y\}_{y=1, ...n}$.
2. The subset of dimensions $\{Dy\}_{y=1, ...n}$ may not be disjoint and they might have diverse cardinalities.
3. For each projected cluster C_y, the projections of the data points in P_y along each dimension in Dy are similar to each other according to a similarity function, but dissimilar to other data points not in C_y.

The first property is based on the fact that relevant dimensions of the clusters contain dense regions in comparison to irrelevant ones and such a concept of "density" is comparatively relative across all the dimensions in the dataset. The reason for the second and third properties is trivial. In the clustering process, which is k-means-based, the Euclidean distance is used in order to measure the similarity between a data point and a cluster center such that only dimensions that contain dense regions are involved in the distance calculation. Hence, the discovered clusters should have, in general, a concave (near spherical) shape.

The algorithm does not assume any distribution on each individual dimension for the input data. Furthermore, there is no restriction imposed on the size of the clusters or the number of relevant dimensions of each cluster. A projected cluster should have a significant number of selected (i.e., relevant) dimensions with high relevance in which a large number of points are close to each other. To achieve this, PCKA proceeds in three phases:

1. *Attribute Relevance Analysis:* The goal is to recognize every dimension in a dataset which demonstrates some cluster structure by discovering dense regions and their location in each dimension. The underlying assumption for this phase is that, in the context of projected clustering, a cluster should have relevant dimensions in which the projection of each point of the cluster is close to a sufficient number of other projected points (from the whole dataset), and this concept of "closeness" is relative across all the dimensions. The identified dimensions represent potential candidates for relevant dimensions of the clusters.
2. *Outlier Handling:* Based on the results of the first phase, the aim is to identify and eliminate outlier points from the dataset. Like the majority of clustering algorithms, PCKA considers outliers as points that do not cluster well.
3. *Discovery of Projected Clusters:* The goal of this phase is to identify clusters and their relevant dimensions. The clustering process is based on a modified version of the k-means algorithm in which the computation of distance is restricted to subsets where the data point values are dense. Based on the identified clusters, in the last step of the algorithm the results of phase 1 are refined by selecting the appropriate dimensions of each cluster.

In PCKA algorithm [20], the objective function defined for clustering is based on k-means algorithm, which groups the data record based on the binary matrix. The assumption of grouping of data record based on the Boolean logic is not much effective. So, we bring the fuzzy concept in order to overcome it using the objective function of FCM clustering. Along with, the time consumption of grouping of record in PCKA is high. So, we bring the gridding concept to reduce the time of clustering process.

3. PCFA: Mining of Projected Clusters in High Dimensional Data using FCM Algorithm

The above discussions stated that clustering of the high-dimensional data is quite a tedious process. In our proposed approach, we intended to develop a clustering method which can handle clustering in high-dimensional data. In accordance with that, we proposed the clustering of data with FCM algorithm. The FCM algorithm that we proposed is a modified one with some associations with binary values. The process is based on a dimensionality reducing algorithm, which helps in reducing the dimensions of the provided dataset. The proposed algorithm is based on four main steps; initially we apply an FCM algorithm to grid the provided dataset. The second step deals with relevant attribute analysis from the gridded dataset, after those outliers are identified and removed from the dataset. The final step deals with the application of modified FCM for clustering the high-dimensional data. The main steps can be pointed out as follows (shown in Figure 1):

1. Gridding.
2. Relevant attribute selection.
3. Outlier detection.
4. Projected clustering with modified-FCM.

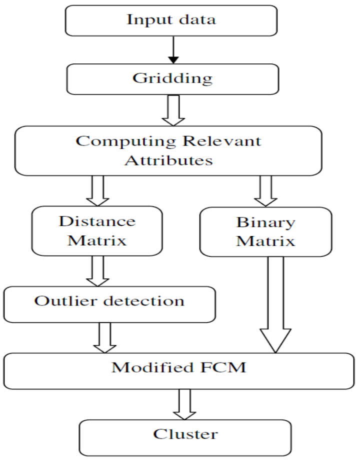


Figure 1. Flow chart of the PCFA algorithm.

1. *Gridding:* The gridding is the initial process in our proposed approach; this strategy is implemented to reduce the time consumption of our approach through searching over each and every data in the dataset. The gridding also, helps in the reducing the space utilization through our proposed method through its particular behavior. In our proposed approach, the gridding is done by FCM algorithm. Initially, we randomly set a number of blocks for the easiness of gridding process. The FCM is applied in the blocks of data, which contains some randomly selected data. A basic FCM algorithm is applied to each block of data with a predefined and fixed number of clusters. The gridding process

replaces the existing data points with a fixed set of reference values. The reference values are obtained from the FCM algorithm. The FCM algorithm selects data that is relevant for one cluster as per the FCM process, after that it produces a centroid value for the cluster. This centroid value serves as the reference values. Similarly centroid value for all the clusters are generated accordingly. The reference value concludes the information of a set data points in the original dataset. Consider the below example:

$$\left|\begin{array}{c:c:c:c} X_1^1, X_1^2, X_1^3, X_1^4 & \cdots & \cdots & \ldots,\ldots, X_1^{d-1}, X_1^d \\ .. & & & . \\ .. & & & . \\ \vdots & & & \ldots,\ldots, X_{n,}^{d-1} X_n^d \end{array}\right| \| FCM \Rightarrow \left|\begin{array}{c:c:c:c} r_1^1 & . & \cdots & .r_1^d \\ . & & & . \\ . & & & . \\ r_m^1 & & & .r_m^d \end{array}\right|$$

In the above example, data points are represented using the *d* and the reference points are represented using *r*. The data points are represented in the left side matrix and the corresponding reference values according to the FCM algorithm are plotted in the right hand side. This reference values are considered for the succeeding processes of our proposed approach.

2. *Computing Relevant Attributes:* Even though the gridding process reduces complexity in terms of time and space, we have to make a relevant attribute analysis for effective clustering process. When we consider a high dimensional data, the clustering algorithm finds quite difficult to process the exact clustering locations. So, it is difficult mark out the relevant attribute to which the cluster belong. We use the projected clustering technique to find out the cluster and the relevant attributes. A projected cluster is a subclass of the data points, while considering the proposed approach; it is a subset of the centroids. Figure below shows a projected cluster model.

In Figure 2, $C=\{c_1,\ldots, c_n\}$, represents the set of centroids and *A1,……, A10* represents the attribute to which the centroids are belonging to, and the projected clusters are shown in blocks. The cluster 1 belongs to attributes A_1, A_5, A_9, so those attributes are considered as the relevant attributes of the cluster 1 and other are considered as the irrelevant attributes. So, our proposed approach has defined a strategy to differentiate the relevant and irrelevant clusters.

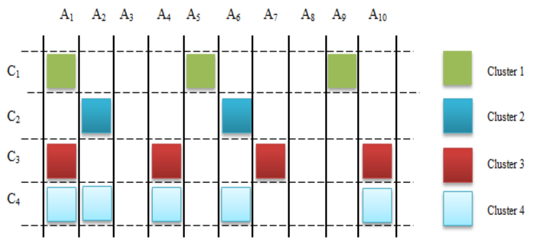


Figure 2. Relevant attribute analysis.

The processing is based on *m×n* matrix. The attributes are arranged in the columns and the centroids are arranged in the rows. In our proposed approach, for finding the attribute analysis a sequential method is adapted. We set up a maximum and minimum value for the reference points in a single column of the matrix, i.e., centroids belonging to a single attribute. The processing is illustrated below:

$$con_C = (max_C^a - min_C^a) \times \alpha \tag{1}$$

$$range_C = \{(C_n - con_c),(C_n + con_c)\} \tag{2}$$

Where, *α* is constant which varies from 0 to 100. We find the difference between the maximum and minimum values and stored in a constant, which is then subjected to find the range for the current set of centroids. The range is calculated is using the above equation which is represented by $range_c$. A count value is selected from the set of centroids which are fall in the value $range_c$. Thus; we find count of all the elements in the *m×n* matrix. With the help of that count values we construct a distance matrix. We define an average function of the reference values in order to find the relevant attribute from the distance matrix:

$$avg_C = {\sum_{i=l}^{n} c_i} \Big/ {N_c}, N_c = No.of\ centroids. \tag{3}$$

The count values are compared with the value generated through the avg_c function and the values which are relevant to the cluster are sorted out and the attribute to which they are related to, are also sorted out. The resultant matrix at this position shows is the details about the relevant attributes, with which we mine the data points. The distance matrix is then converted into a binary matrix by comparing with the avg_c value. The values which fall over the avg_c will be replaced by 1 and the rest with 0. Thus, the resultant will be a same dimensioned binary matrix instead of the distance matrix.

3. *Outlier Detection:* The outlier detection is a process, which removes the non-relevant data points from the resultant matrix, which is obtained from the distance matrix. The outlier detection is done with a row wise operation on the attributes. The row wise operation gives a distance matrix as output. It is a sorted matrix. In this step also we define an average function for the attributes:

$$avg_A = \frac{\sum_{i=l}^{m} A_i}{M_A}, M_A = No.\ of\ \ Attributes \tag{4}$$

We convert the distance matrix into sorted distance matrix by comparing the count values with the value generated according to avg_A. We sort the attribute list according to the value generated according to the comparison process. Then, a Minimum and maximum value are generated to find the outliers from the distance matrix:

$$max = A_{2nd\ lowest}, min = A_{lowest} \tag{5}$$

The max value is set as the attribute with second lowest value and the min values is set as the attribute with a lowest value. A function is set to calculate the outlier from the distance matrix with the help if the *max* and *min* values:

$$Con_A = (max - min) * 0.3 \tag{6}$$

The value of the function Con_A serves as threshold; the values of the elements which are less than the threshold are expelled from the binary matrix as outliers. Since the distance matrix is a sorted matrix, the outliers are which are expelled is proportional to its original positions. If a number of elements which belong to a single attribute are found as outliers then that whole column is expelled as outliers. After the outlier detection process, the elements are subjected to the clustering process.

4. *Projected Clustering with Modified FCM:* This step includes the clustering of the reference values, which represents the value of data, points from the dataset. In our proposed approach, we consider the reference values as the input to the FCM algorithm. The binary matrix helps in reducing the high dimensionality of the actual data set and the binary values in the binary matrix helps in finding membership functions accurately. The processing can be illustrated as follows.

Initially, the number of cluster and the reference values are given as inputs. According to the proposed approach, the following objective function has to be minimized:

$$F_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij}^{z} \; \| X_i^{ref} - C_j^{mod} \|^{2} \tag{7}$$

Where *m, n* are any real number greater than 1, u_{ij} is the degree of membership of in the cluster j, C_i^{ref} is the i^{th} of d-dimensional reference values, C_j^{mod} is the d-dimension center of the cluster. The objective function is iteratively optimized to get the expected result. The Updation of the membership values and the cluster centroids helps in the optimization process. The membership function can be updated using the following expression:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| (X_i^{ref} - C_j^{mod}) \times b_i \right\|}{\left\| (X_i^{ref} - C_k^{mod}) \times b_i \right\|} \right)^{\frac{2}{m-1}}} \tag{8}$$

Where, k is the iteration step and the b_i is the binary values from the binary matrix, which is corresponding to the current reference value. Similarly, the new centroid is also, updated with the following expression:

$$C_j^{mod} = \frac{\sum_{i=1}^{n} u_{ij}^{z} . X_i^{ref}}{\sum_{i=1}^{n} u_{ij}^{z}} \tag{9}$$

The modifications, which we are incorporated in the traditional FCM algorithm improves the efficiency of the normal FCM algorithm in the case of projected mining of the clusters. The binary values from the binary matrix the root course of the improvement in the traditional FCM algorithm.

4. Results and Discussion

The results obtained from the experimentation of the proposed hybrid clustering algorithm with different datasets are presented in this section. We have implemented the proposed hybrid clustering system using Java (jdk 1.6). We have experimented with two different datasets obtained from Infobiotics Protein Structure Prediction (PSP) benchmarks repository [10] and UCI machine learning repository [9].

4.1. Experimental Set Up and Datasets

This experimental environment of proposed PCFA algorithm for high-dimensional data is experimented with Windows 7 Operating system at a 2.1 GHz core i5 PC machine with 4 GB main memory running a 64 bit version of Windows. The experimental analysis of the proposed research methodology is presented in this section. For extensive analysis, we have utilized the very large high dimensional dataset taken from the Infobiotics PSP benchmarks repository.

- *Dataset Description:* The datasets we have utilized here for evaluation purposes is taken from Infobiotics PSP benchmarks repository [10] and UCI machine learning repository [9].
  a. *Dataset 1 (DB1):* This dataset is taken from Infobiotics PSP benchmarks repository [10] that contains an adjustable real-world family of benchmarks suitable for testing the scalability of classification/regression methods. The dataset DB1 taken for experimentation contains data with 40 attributes and 100000 records.
  b. *Dataset 2 (DB2):* This dataset, Census-Income (KDD) dataset, is taken from the UCI machine learning repository [10] and it contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the US Census Bureau. The dataset DB2 taken for experimentation contains data with 40 attributes and 100000 records.

4.2. Evaluation Metrics

The obtained results from the proposed research are analyzed with the aid of the evaluation metrics such as clustering accuracy, memory usage and the computational time. We have used the clustering accuracy described in [27], for evaluating the performance of the proposed approach. The evaluation metric used in the proposed approach is given below, Clustering Accuracy (CA):

$$CA = \frac{1}{N}\sum_{i=1}^{T} X_i^{ref} \qquad (10)$$

Where, N → Number of data points in the reference value dataset:

T →Number of resultant cluster.

X_i → Number of data points occurring in both cluster i and its corresponding class.

Computational time indicates the time taken to execute the computer program and it typically grows with the input size and the Memory usage defines the memory utilized by the current jobs present in the particular system.

4.3. Performance Evaluation

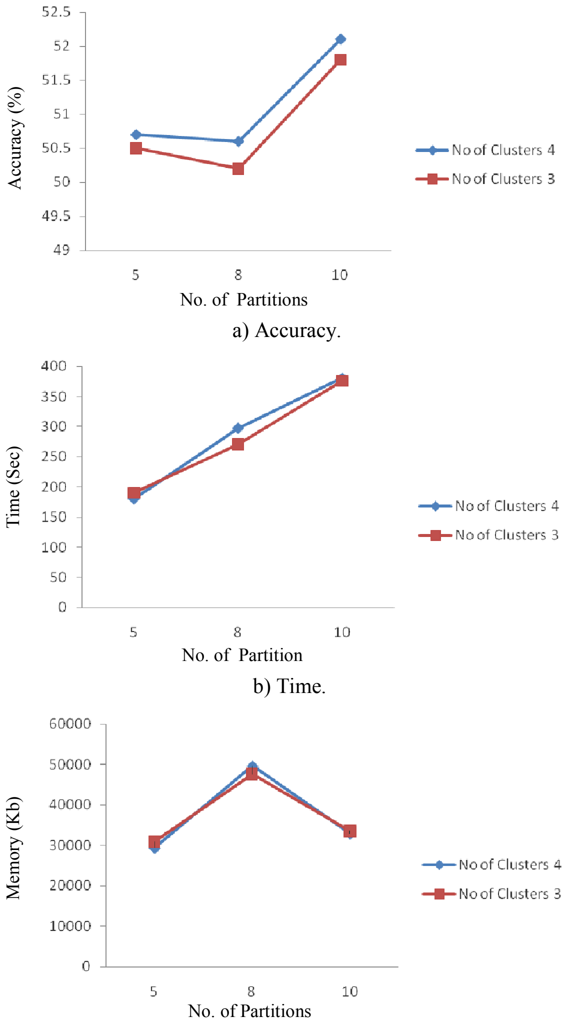
The performance of the proposed approach has been evaluated in terms of the evaluation metrics, memory usage and the computational time. The metrics, memory usage and the time have been analyzed with the two different high dimensional datasets by giving different *k* values and the thresholds. The evaluation results, clustering accuracy and the computational time of the proposed approach are plotted as a graph and shown in the following Figures.

4.3.1. Performance Analysis of Proposed Approach on DB1

The performance analysis results of the proposed approach on dataset 1 are presented in this section. In the following sections, we plot the accuracy, memory usage and execution time of the proposed PCFA algorithm. The results are processed by varying the size of no. of partitions and the Alpha value (α). The graphs are plotted according to the variation in number of partitions, alpha variation and cluster size.

a. *Effect of Number of Partitions:* The effect of number of partitions on accuracy of the proposed PCFA algorithm is plotted in Figure 3. From Figure 3, we can see the accuracy tends to a high level at partition value 10 and for the rest of the values it produces almost similar accuracy values. The execution time is in a proportional manner, i.e., as the no. of partitions increases, the execution time also increases. On the other hand, the case is entirely different for the memory usage. In that scenario, there is no specific variation whenever the final cluster number varied.

b. *Effect of Alpha Variations on the Proposed Approach:* Figure 4 illustrates the effect of alpha variations, *α* on the accuracy of the proposed PCFA algorithm. Figure depicts the level of accuracy to reach feasible value at alpha range 30. The accuracy values are approximately similar for the remainder values. When the range of alpha mounts, there is a corresponding increase in the execution time also, this means that the execution time is following a proportional mode. If we consider the scenario of memory usage, the case would not be exactly the same. In that case, if we change the count of final cluster, there will not be any explicit variation.



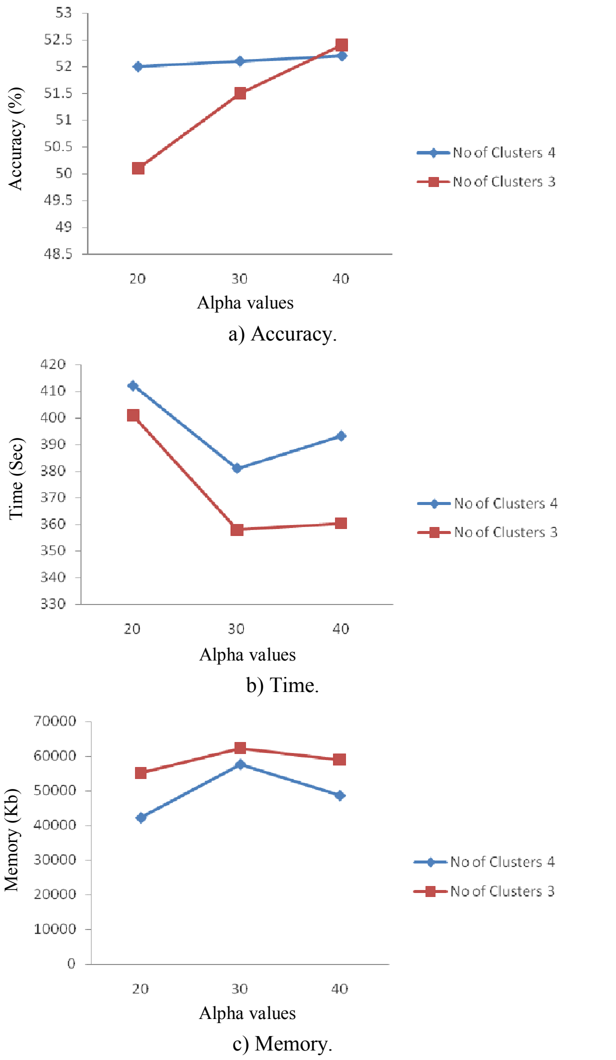
a) Accuracy.

b) Time.

c) Memory.

Figure 3. Effects of proposed algorithm based on different no. of partitions.

c. *Effect of Alpha Variations on The Proposed Approach:* Figure 4 illustrates the effect of alpha variations, *α* on the accuracy of the proposed PCFA algorithm. Figure depicts the level of accuracy to reach feasible value at alpha range 30. The accuracy values are approximately similar for the remainder values. When the range of alpha mounts, there is a corresponding increase in the execution time also, this means that the execution time is following a proportional mode. If we consider the scenario of memory usage, the case would not be exactly the same. In that case, if we change the count of final cluster, there will not be any explicit variation.



a) Accuracy.

b) Time.

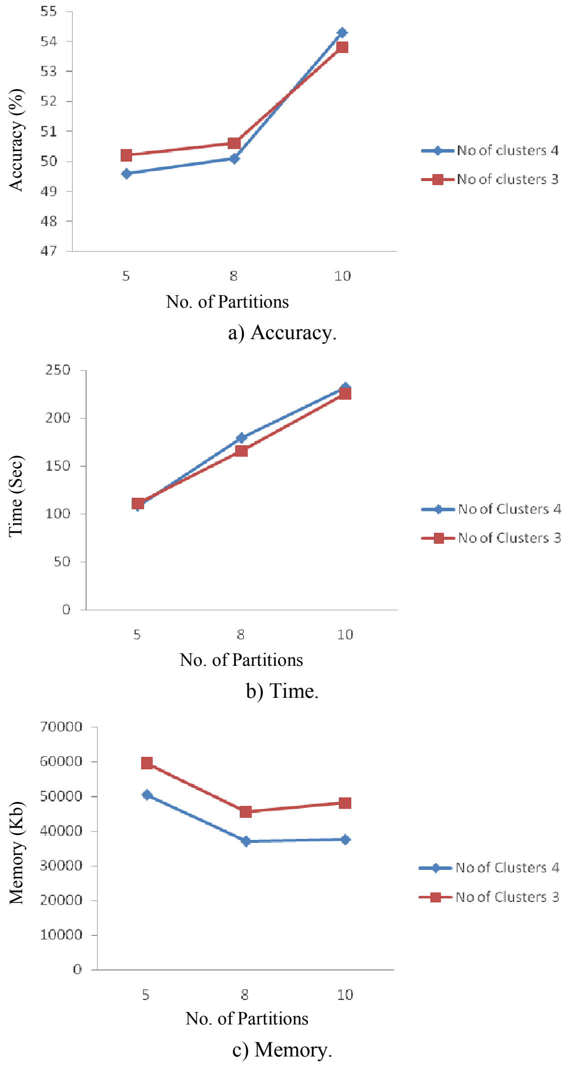
c) Memory.

Figure 4. Effects of proposed algorithm based on different alpha values.

4.3.2. Performance Analysis of Proposed Approach on DB2

The DB2 also checked under the parameters, no. of partitions and alpha values. The responses of the proposed PCFA algorithm is plotted in the below section. The graphs are plotted according to the variation in number of partitions, alpha variation and cluster size.

a. *Effect of Number of Partitions:* The effect of number of partitions on accuracy is analyzed with the help of Figure 5, in which we can see that, the accuracy tends to a high level at partition value 10 and for the rest of the values it produces almost similar accuracy values. The execution time is in a proportional manner as in DB1, i.e., as the No. of partitions increases, the execution time also increases. When we consider the extent of the memory usage on different no. of partitions, it shows a sudden increase in the lower and higher number partitions, but it is less affine when the partition is set to 8. From Figure 5, we can conclude that if we give more partitions, accuracy is increased from 50 to 54% and at the same time, the time and memory consumption is also less.



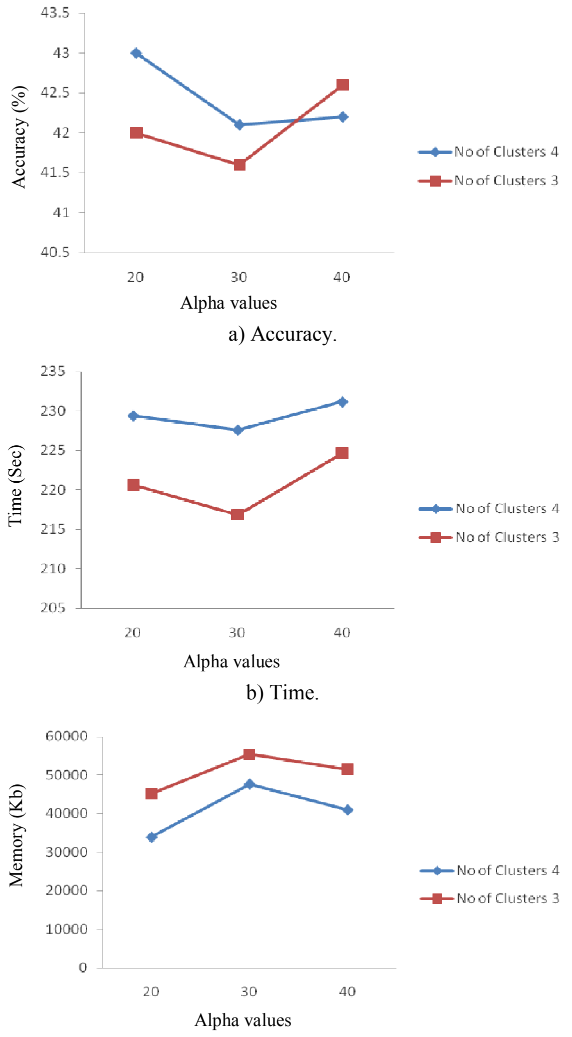
a) Accuracy.

b) Time.

c) Memory.

Figure 5. Effects of proposed algorithm based on different no. of partitions.

b. *Effect of Alpha variations on the proposed approach:* In the following section we illustrate the effect of alpha values on accuracy, execution time and memory usage. Figure a, stands for the evaluation of accuracy based on the different alpha values. The accuracy value tends to a feasible value at alpha range 20. Figures b and c, represents the performance of DB2 based on execution time and memory usage for different alpha values. Figure b and c, states that, as the alpha value increases the values of memory usage and execution time decreases. Thus, we can finalize that, the PCFA algorithm is feasible at alpha range 30. From Figure 6, we observe that fixing an appropriate alpha value is important to maintain accuracy in a good way. Meanwhile, time and memory consumption should be low. For lower alpha values, accuracy seems good compared with higher alpha values. If alpha value is 30, time and memory consumption is less compared with 20 and 40%.



a) Accuracy.

b) Time.

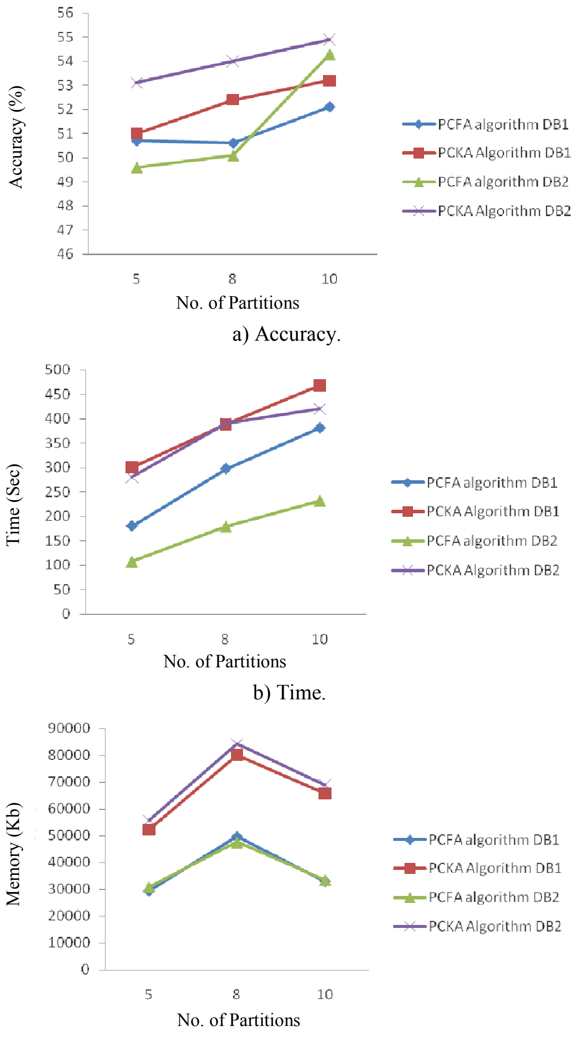
c) Memory.

Figure 6. Effects of proposed algorithm based on different alpha values.

4.4. Comparative Analysis

The comparative analysis deals with the comparison of proposed PCFA algorithm and the PCKA algorithm [20]. The comparison graphs are plotted in Figure 7. In the comparison analysis, we check the accuracy, execution time and memory usage of the dataset DB1 and DB2 with both the above mentioned methods. The comparison analysis shows that our proposed approach has advantage over the PCKA algorithm in the case of execution time and memory space usage. In term of accuracy, the PCFA algorithm is not much good because the gridded data is used for clustering. But, time and memory consumption is less for the PCFA algorithm compared with the PCKA algorithm.

The proposed PCFA algorithm is then compared with the traditional FCM algorithm. The comparison of the traditional FCM algorithm with our PCFA shows the improved performance of the PCFA algorithm, which states that the modification in the FCM has improved the performance of the normal FCM algorithm.



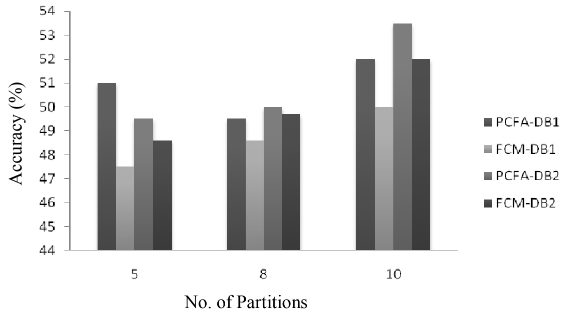
a) Accuracy.

b) Time.

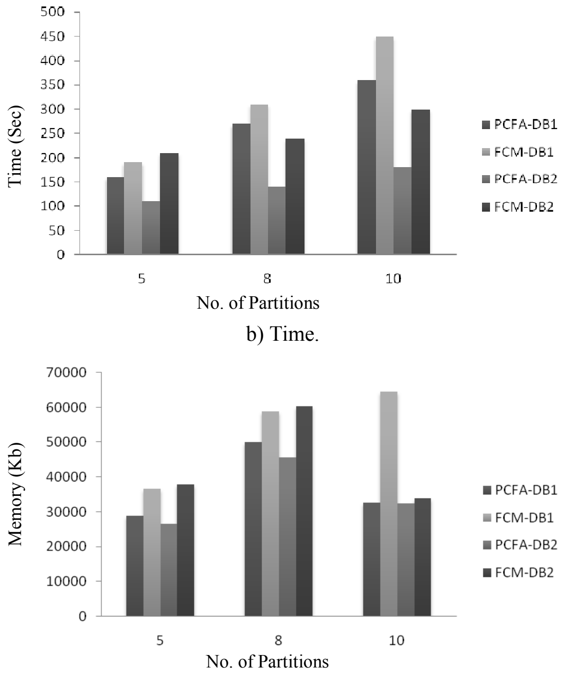
c) Memory.

Figure 7. Comparison analysis.

The improvement in the proposed modified FCM algorithm can be visible from Figure 8. It shows that the modified FCM has the better accuracy as compared to the conventional FCM algorithm. The time for execution and memory usage is very less as compared to the normal FCM algorithm. Thus, we can state that the modification made in the FCM has improved the conventional FCM algorithm.



a) Accuracy.



b) Time.

c) Memory.

Figure 8. Comparison analysis of PCFA against FCM.

5. Conclusions

The latest researches are become the catalysts of proposing a new approach for clustering high-dimensional data. Here, we have proposed an algorithm, called PCFA. Initially, the standard FCM is used for initial processing called gridding. The processes after gridding is the relevant attribute detection and the outlier detection. The data after preparatory process are subjected for the processing with modified FCM algorithm. The proposed approach performs well under different test criteria. The result of the proposed approach shows the upper hand of our proposed approach over the existing PCKA approach. Comparing to the previous method, our proposed approach has less execution time and memory usage. The evaluation results showed that, the PCFA algorithm shows approximately 20% improvement in the execution time and 50% improvement in memory usage over the PCKA algorithm.

References

[1] Aggarwal C. and Yu P., “Finding Generalized Projected Clusters in High Dimensional Spaces,” *in Proceedings of ACM SIGMOD International Conference Management of Data*, New York, USA, pp. 70-81, 2000.

[2] Aggarwal C., “A Human-Computer Interactive Method for Projected Clustering,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 448-460, 2004.

[3] Aggarwal C., Wolf J., Yu P., Procopiuc C., and Park J., “Fast Algorithms for Projected Clustering,” *in Proceedings of the ACM SIGMOD International Conference on Management of Data*, New York, UAS, pp. 61-72, 1999.

[4] Agrawal R. and Srikant R., “Fast Algorithms for Mining Association Rules in Large Databases,” *in Proceedings of the 20th International Conference on Very Large Databases*, San Francisco, USA, pp. 487-499, 1994.

[5] Agrawal R., Mannila H., Srikant R., Toivonen H., and Verkamo A., “Fast Discovery of Association Rules,” *Advances in Knowledge Discovery and Data Mining*, pp. 307-328, 1996.

[6] Bezdek J., *Pattern Recognition with Fuzzy Objective Function Algoritms*, Plenum Press, New York, 1981.

[7] Bharat G. and Durga T., “A System for Outlier Detection of High Dimensional Data,” *the International Journal of Computer Science & Informatics*, vol. 1, no. 2, pp. 97-101, 2011.

[8] Bharat P. and Durga T., “Hierarchical Clustering of Projected Data Streams using Cluster Validity Index,” *Advances in Computer Science and Information Technology*, vol. 131, no. 4, pp. 551-559, 2011.

[9] Census income (KDD) Datasets from UCI Machine Learning Repository, available at: http://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD), last visited 2000.

[10] Datasets from Infobiotics PSP (Protein Structure Prediction) benchmarks repository, available at: http://icos.cs.nott.ac.uk/datasets/psp_benchmark.html, last visited 2008.

[11] Dunn J., “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters,” *Journal of Cybernetics*, vol. 3, no. 3, pp. 32-57, 1973.

[12] Farnstrom F., Lewis J., and Elkan C., “Scalability for Clustering Algorithms Revisited,” *ACM SIGKDD Explorations Newsletter*, vol. 2, no.1, pp. 51-57, 2000.

[13] Gabriela M. and Jorg S., “Finding Non-Redundant, Statistically Significant Regions in High Dimensional Data: A Novel Approach to Projected and Subspace Clustering,” *in Proceedings of the 14th ACM SIGKDD International Conference Knowledge Discovery and Data Mining*, New York, USA, pp. 533-541, 2008.

[14] Gabriela M., Arthur Z., Peer K., Hans K., and Jörg S., “Subspace and Projected Clustering: Experimental Evaluation and Analysis,” *Knowledge and Information Systems*, vol. 21, no. 3, pp. 299-326, 2009.

[15] Gnanabaskaran A. and Duraiswamy K., “An Efficient Approach to Cluster High Dimensional Spatial Data Using K-Mediods Algorithm,” *European Journal of Scientific Research*, vol. 49 no. 4, pp. 617-624, 2011.

[16] Jhimli A., Pralhad R., and Animesh A., "Clustering Items in Different Data Sources Induced by Stability," *the International Arab Journal of Information Technology*, vol. 6, no. 4, pp. 394-402, 2009.

[17] Kevin Y., David W., and Michael K., "A Highly-Usable Projected Clustering Algorithm for Gene Expression Profiles," *in Proceedings of the 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, Washington, USA, pp. 41-48, 2003.

[18] Lance P., Ehtesham H., and Huan L., "Subspace Clustering for High Dimensional Data: A Review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90-105, 2004.

[19] Man L. and Nikos M., "Iterative Projected Clustering by Subspace Mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 176-189, 2005.

[20] Mohamed B. and Shergrui W., "Mining Projected Clusters in High-Dimensional Spaces," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 4, pp. 507-522, 2009.

[21] Mugunthadevi K., Punitha S., and Punithavalli M., "Survey on Feature Selection in Document Clustering," *the International Journal on Computer Science and Engineering*, vol. 3, no. 3, pp. 1240-1244, 2011.

[22] Naveen K., Naveen G., and Veera R., "Partition Algorithms-A Study and Emergence of Mining Projected Clusters in High-Dimensional Dataset," *the International Journal of Computer Science and Telecommunications*, vol. 2, no. 4, pp. 34-37, 2011.

[23] Sembiring R., Zain J., and Abdullah E., "Clustering High Dimensional Data using Subspace and Projected Clustering Algorithms," *the International Journal of Computer Science & Information Technology*, vol. 2, no. 4, pp. 162-170, 2010.

[24] Singh V., Sahoo L., and Kelkar A., "Mining Subspace Clusters in High Dimensional Data," *the International Journal of Recent Trends in Engineering and Technology*, vol. 3, no. 1, pp. 118-112, 2010.

[25] Yeung K. and Ruzzo W., "An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data," *Bioinformatics*, vol. 17, no. 9, pp. 763-774, 2001.

[26] Yip K. and Cheung D., "HARP: A Practical Projected Clustering Algorithm," *IEEE Transactions Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1387-1397, 2004.

[27] Zhexue H., "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283-304, 1998.

Ilango Murugappan obtained his MCA degree in computer applications from Alagappa University and perusing his PhD degree in Anna University, Chennai in clustering domain. Since 1999, he is working in KLN College of Engineering, India. His research area of interest: DBMS, data mining and data warehousing.

Mohan Vasudev obtained his post graduate degree in Mathematics from Mysore University in 1973 and PhD degree from IIT Bombay in 1978. Since 1978, he is working in Thiagarajar College of Engineering, Madurai. His research area of interest: computer applications, computational methods, fluid mechanical and simulation.