MACHINE LEARNING

ASSIGNMENT – 3

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is an application of clustering?

a. Biological network analysis

b. Market trend prediction

c. Topic modeling

d. All of the above

answer: b. Market trend prediction

2. On which data type, we cannot perform cluster analysis?

a. Time series data

b. Text data

c. Multimedia data

d. None

answer: a. Time series data

3. Netflix's movie recommendation system uses

a. Supervised learning

b. Unsupervised learning

c. Reinforcement learning and Unsupervised learning

d. All of the above

4. The final output of Hierarchical clustering is

a. The number of cluster centroids

b. The tree representing how close the data points are to each other

c. A map defining the similar data points into individual groups

d. All of the above

5. Which of the step is not required for K-means clustering?

a. A distance metric

b. Initial number of clusters

c. Initial guess as to cluster centroids

d. None

6. Which is the following is wrong?

a. k-means clustering is a vector quantization method

b. k-means clustering tries to group n observations into k clusters

c. k-nearest neighbour is same as k-means

d. None

answer: . k-means clustering is a vector quantization method

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

i. Single-link

ii. Complete-link

iii. Average-link

Options:

a.1 and 2

b. 1 and 3

c. 2 and 3

d. 1, 2 and 3

answer: d. 1, 2 and 3

8. Which of the following are true?

i. Clustering analysis is negatively affected by multicollinearity of features

ii. Clustering analysis is negatively affected by heteroscedasticity

Options:

a. 1 only

b. 2 only

c. 1 and 2

d. None of them

answer: a. 1 only

MACHINE LEARNING

ASSIGNMENT – 3

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

a. 2

b. 4

c. 3

d. 5

answer: a. 2

10. For which of the following tasks might clustering be a suitable approach?

a.Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

b.Given a database of information about your users, automatically group them into different market segments.

c. Predicting whether stock price of a company will increase tomorrow.

d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

Answer: a.Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

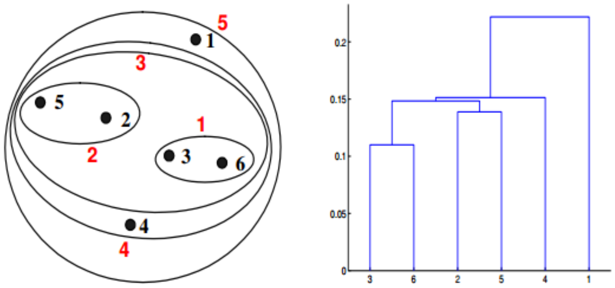
11. Given, six points with the following attributes:

MACHINE LEARNING

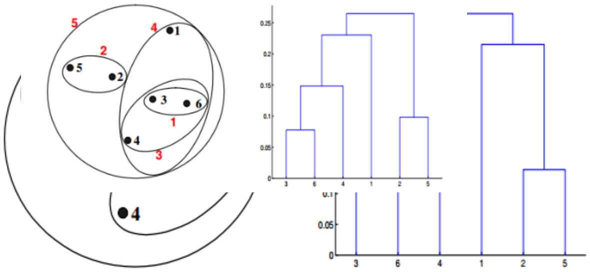
ASSIGNMENT – 3

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

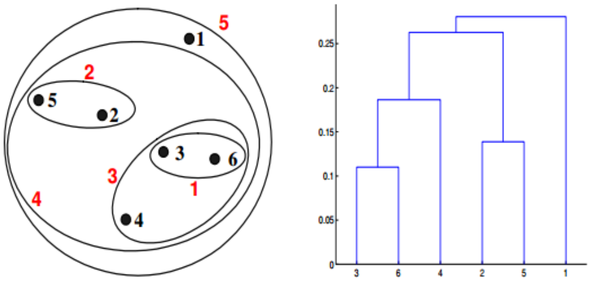
a.



b



c.



d.

Answer:a

MACHINE LEARNING

ASSIGNMENT – 3

12. Given, six points with the following attributes:

Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

a.

b.

MACHINE LEARNING

ASSIGNMENT – 3

c.

d.

Answer: b

For the single link or MAX version of hierarchical clustering, the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters. Similarly, here points 3 and 6 are merged first. However, {3, 6} is merged with {4}, instead of {2, 5}. This is because the dist({3, 6}, {4}) = max(dist(3, 4), dist(6, 4)) = max(0.1513, 0.2216) = 0.2216, which is smaller than dist({3, 6}, {2, 5}) = max(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = max(0.1483, 0.2540, 0.2843, 0.3921) = 0.3921 and dist({3, 6}, {1}) = max(dist(3, 1), dist(6, 1)) = max(0.2218, 0.2347) = 0.2347.

Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly

13. What is the importance of clustering?

Data professionals often use clustering in the Exploratory Data Analysis phase to discover new information and patterns in the data. As clustering is unsupervised machine learning, it doesn't require a labeled dataset. Clustering itself is not one specific algorithm but the general task to be solved.

Importance — Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships.

The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the —better‖ or more distinct the clustering. Data mining is the process of analysing data from different viewpoints and summerising it into useful information.

Data mining is one of the top research areas in recent days. Cluster analysis in data mining is an important research field it has its own unique position in a large number of data analysis and processing.

I. Introduction Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

II. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering. In this paper, clustering analysis is done. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is important in data analysis and data mining applications[1]. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. A good clustering algorithm is able to identity clusters irrespective of their shapes. The stages involved in clustering algorithm are as follows,

III. Literature Review Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results.

IV. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to Raw data clustering algorithms clusters of data International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016 ISSN 2229-5518 247 IJSER © 2016 http://www.ijser.org IJSER modify data preprocessing and model parameters until the result achieves the desired properties…..

V. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goal. Why clustering? • Organizing data into clusters shows internal structure of the data – Ex. Clusty and clustering genes above • Sometimes the partitioning is the goal – Ex. Market segmentation • Prepare for other AI techniques – Ex. Summarize news (cluster and then find centroid) • Techniques for clustering is useful in knowledge discovery in data –

VI. Ex. Underlying rules, reoccurring patterns, topics, etc. Methods: Basic Agglomerative Hierarchical Clustering Algorithm 1) Compute the proximity graph, if necessary. (Sometimes the proximity graph is all that is available.) 2) Merge the closest (most similar) two clusters. 3) Update the proximity matrix to reflect the proximity between the new cluster and the original clusters. 4) Repeat steps 3 and 4 until only a single cluster remains. The key step of the previous algorithm is the calculation of the proximity between two clusters, and this is where the various agglomerative hierarchical techniques differ.

VII. Any of the cluster proximities that we discuss in this section can be viewed as a choice of different parameters (in the Lance-Williams formula) for the proximity between clusters Q and R, where R is formed by merging clusters A and B. p(R, Q) = ⟨A p(A, Q) + ⟨B p(B, Q) + ⟨⟨p(A, Q) + ⟨⟨| p(A, Q) – p(B, Q) | In words, this formula says that after you merge clusters A and B to form cluster R, then the distance of the new cluster, R, to an existing cluster, Q, is a linear function of the distances of Q from the original clusters A and B. Any hierarchical technique that can be phrased in this way does not need the original points, only the proximity matrix, which is updated as clustering occurs. However, while a general formula is nice, it is often easier to understand the different hierarchical methods by looking directly at the definition of cluster distance that each method uses, and that is the approach that we shall take here. [DJ88] and [KR90] both give a table that describes each method in terms of the Lance-Williams formula .

VIII. Mutual Nearest Neighbor Clustering Mutual nearest neighbor clustering is described in [GK77]. It is based on the idea of the ―mutual neighborhood value (mnv)‖ of two points, which is the sum of the ranks of the two points in each other's sorted nearest-neighbor lists. Two points are then said to be mutual nearest neighbors if they are the closest pair of points with that mnv. Clusters are built up by starting with points as singleton clusters and then merging the closest pair of clusters, where close is defined in terms of the mnv. The mnv between two clusters is the maximum mnv between any pair of points in the combined cluster. If there are ties in mnv between pairs of clusters, they are resolved by looking at the original distances between points.

IX. Thus, the algorithm for mutual nearest neighbor clustering works in the following way.

X. a) First the k-nearest neighbors of all points are found. In graph terms this can be regarded as breaking all but the k strongest links from a point to other points in the proximity graph.

XI. b) For each of the k points in a particular point's knearest neighbor list, calculate the mnv value for the two points. It can happen that a point is in one point's knearest neighbor list, but not vice-versa. In that case, set the mnv value to some value larger than 2k.

XII. c) Merge the pair of clusters having the lowest mnv (and the lowest distance in case of ties).

XIII. d) Repeat step (c) until the desired number of clusters is reached or until the only clusters remaining cannot be merged.

XIV. The latter case will occur when no points in different clusters are k-nearest neighbors of each other. The mutual nearest neighbor technique has behavior similar to the shared nearest neighbor technique in that it can handle clusters of varying density, size, and shape. However, it is basically hierarchical in nature while the shared nearest neighbor approach is partitional in nature. Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance.

XV. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

XVI. Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance to) to use.

XVII. Popular choices are known as

XVIII. single-linkage clustering (the minimum of object distances),

XIX. complete linkage clustering (the maximum of object distances) or UPGMA ("Unweighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering).

XX. Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions). These methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (known as "chaining phenomenon", in particular with single-linkage clustering). In the general case, the complexity is for agglomerative clustering and for divisive clustering, which makes them too slow for large data sets.

XXI. For some special cases, optimal efficient methods (of complexity) are known:

XXII. SLINK for single-linkage and CLINK for completelinkage clustering. In the data mining community these methods are recognized as a theoretical foundation of cluster analysis, but often considered obsolete. They did however provide inspiration for many later methods such as density based clustering.

XXIII. • Linkage clustering examples • Single-linkage on Gaussian data. At 35 clusters, the biggest cluster starts fragmenting into smaller parts, while before it was still connected to the second largest due to the single-link effect.

XXIV. • Single-linkage on density-based clusters. 20 clusters extracted, most of which contain single elements, since linkage clustering does not have a notion of "noise". International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016 ISSN 2229-5518 249 IJSER © 2016 http://www.ijser.org IJSER

XXV. General Types of Clusters 1. Well-separated clusters A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in the cluster.

XXVI. 2. Centre-based clusters A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the —centre‖ of a cluster, than to the centre of any other cluster [2]. The centre of a cluster is often a centroid. 3. Contiguous clusters A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster. 4. Density-based clusters A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density. 5. Shared Property or Conceptual Clusters Finds clusters that share some common property or represent a particular concept. III.

Analysis of Clustering Algorithm Clustering is the main task of Data Mining and it is done by the number of algorithms.

XXVII. The most commonly used algorithms in Clustering are Hierarchical, Partitioning and Grid based algorithms

XXVIII. 1. Hierarchical Algorithms Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It is the connectivity based clustering algorithms. The hierarchical algorithms build clusters gradually. Hierarchical clustering generally fall into two types: In hierarchical clustering, in single step, the data are not partitioned into a particular cluster [3].

XXIX. **IV. The Application of Cluster Analysis in Data Mining The application of cluster analysis in** data mining has two main aspects: first, clustering analysis can be used as a pre-processing step for the other algorithms such as features and classification algorithm, and also can be used for further correlation analysis.

XXX. Second, it can be used as a stand-alone tool in order to get the data distribution, to observe each cluster features, then focus on a specific cluster for some further analysis [5]. Cluster analysis can be available in market segmentation, target customer orientation, performance assessment, biological species etc.

XXXI. **Some Relative Applications The cluster analysis has been applied to many occasions. For** example, in commercial, cluster analysis was used to find the different customer groups, and summarize different customer group characteristics through the buying habits; in biotechnology, cluster analysis was used to categorized animal and plant populations according to population and to obtain the latent structure of knowledge; in geography, clustering can help biologists to determinate the relationship of the different species and different geographical climate; in the banking sector, by using cluster analysis to bank customers to refine a user group; in the insurance industry, according to the type of residence, around the business district, the geographical location, cluster analysis can be used to complete an automatic grouping of regional real estate, to reduce the manpower cost and insurance company industry risk; in the Internet, cluster analysis was used for document classification and information retrieval etc.

XXXII. **The overall goal of the data mining process is to extract information from a large data set** and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can be done by the different no. of algorithms such as hierarchical, partitioning and grid algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering; the value of k-mean is set. The application of cluster analysis is more and more urgent; the requirements are also getting higher and higher. With the development of modern technology, in the near future, cluster areas will achieve a critical breakthrough.

14. How can I improve my clustering performance?

Answer: Clustering analysis is one of the main analytical methods in data mining. K-means is the most popular and partition based clustering algorithm. But it is computationally expensive and the quality of resulting clusters heavily depends on the selection of initial centroid and the dimension of the data.

Several methods have been proposed in the literature for improving performance of the k-means clustering algorithm.

Principal Component Analysis (PCA) is an important approach to unsupervised dimensionality reduction technique. This paper proposed a method to make the algorithm more effective and efficient by using PCA and modified k-means.

In this , we have used Principal Component Analysis as a first phase to find the initial centroid for k-means and for dimension reduction and k-means method is modified by using heuristics approach to reduce the number of distance calculation to assign the data-point to cluster.

By comparing the results of original and new approach, it was found that the results obtained are more effective, easy to understand and above all, the time taken to process the data was substantially reduced.

1. Data mining is a convenient way of extracting patterns, which represents knowledge implicitly stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability.
2. It can be viewed as an essential step in the process of knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared for the mining task.
3. Data mining can be performed on various types of databases and information repositories, but the kind of patterns to be found are specified by various data mining functionalities like class description, association, correlation analysis, classification, prediction, cluster analysis etc. Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets.
4. It is a process of grouping data objects into disjoint clusters so that the International Journal of Database Management Systems (IJDMS), Vol.3, No.1, February 2011 197 data in the same cluster are similar, and data belonging to different cluster are differ. Many algorithms have been developed for clustering. A clustering algorithm typically considers all features of the data in an attempt to learn as much as possible about the objects. However, with high dimensional data, many features are redundant or irrelevant. The redundant features are of no help for clustering; even worse, the irrelevant features may hurt the clustering results by hiding clusters in noises. There are many approaches to address this problem.
5. The simplest approach is dimension reduction techniques including principal component analysis (PCA) and random projection. In these methods, dimension reduction is carried out as a preprocessing step. K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast,
6. so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. The standard k-means algorithm [10, 14] is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high in high dimensional data.
7. Different methods have been proposed [4] by combining PCA with k-means for high dimensional data. But the accuracy of the k-means clusters heavily depending on the random choice of initial

centroids. If the initial partitions are not chosen carefully, the computation will run the chance of converging to a local minimum rather than the global minimum solution. The initialization step is therefore very important.

8. To combat this problem it might be a good idea to run the algorithm several times with different initializations. If the results converge to the same partition then it is likely that a global minimum has been reached. This, however, has the drawback of being very time consuming and computationally expensive.
9. In this work, initial centers are determined using PCA and k-means method is modified by using heuristic approach to assign the data-point to cluster. This paper deals with the method for improving the accuracy and efficiency by reducing dimension and initialize the cluster for modified k-means using PCA.

2. **K-MEANS CLUSTERING ALGORITHM K-means is a prototype-based, simple partitional clustering** technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids.

A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice.

The algorithm consist of two separate phases: the first phase is to select k centers randomly, where the value of k is fixed in advance. The next phase is to assign each data object to the nearest center. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects are included in some clusters, recalculating the average of the clusters.

This iterative process continues repeatedly until the criterion function becomes minimum. The kmeans algorithm works as follows: a) Randomly select k data object from dataset D as initial cluster centers. b) Repeat a. Calculate the distance between each data object di(1<=i<=n) and all k cluster centers cj(1<=j<=n) and assign data object di to the nearest cluster. b. For each cluster j(1<=j<=k), recalculate the cluster center. c. Until no changing in the center of clusters.

International Journal of Database Management Systems (IJDMS), Vol.3, No.1, February 2011 198 The most widely used convergence criteria for the k-means algorithm is minimizing the SSE. 2 k SSE xi j j 1x c i j = − ∑ ∑ μ = ∈ Where 1 μ x j i n x c j i j = ∑ ∈ Denotes the mean of cluster cj and nj denotes the no. of instances in cj .

The k-means algorithm always converges to a local minimum. The particular local minimum found depends on the starting cluster centroids. The k-means algorithm updates cluster centroids till local minimum is found. Before the k-means algorithm converges, distance and centroid calculations are done while loops are executed a number of times, say l, where the positive integer l is known as the number of k-means iterations.

The precise value of l varies depending on the initial starting cluster centroids even on the same dataset. So the computational complexity of the algorithm is O(nkl), where n is the total number of objects in the dataset, k is the required number of clusters and l is the number of iterations. The time complexity for the high dimensional data set is O(nmkl) where m is the number of dimensions.

3. PRINCIPAL COMPONENT ANALYSIS As a preprocessing stage of data mining and machine learning, dimension reduction not only decreases computational complexity, but also significantly improves the accuracy of the learned models from large data sets.

PCA [11] is a classical multivariate data analysis method that is useful in linear feature extraction. Without class labels it can compress the most information in the original data space into a few new features, i.e., principal components. Handling high dimensional data using clustering techniques obviously a difficult task in terms of higher number of variables involved.

In order to improve the efficiency, the noisy and outlier data may be removed and minimize the execution time and we have to reduce the no. of variables in the original data set The central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables.

It is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set. Principal Component A data set xi (i= 1,…,n) is summarized as a linear combination of ortho-normal vectors called principal components, which is shown in the Figure 1.

The steps involved in PCA are

Step1: Obtain the input matrix Table

Step2: Subtract the mean

Step3: Calculate the covariance matrix

Step4: Calculate the eigenvectors and eigenvalues of the covariance matrix

Step5: Choosing components and forming a feature vector Step6: deriving the new data set.

The eigenvectors with the highest eigenvalue is the principal component of the data set. In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. To reduce the dimensions, the first d (no. of principal components) eigenvectors are selected. The final data has only d dimensions.

The main objective of applying PCA on original data before clustering is to obtain accurate results so that the researchers can do analysis in better way. Secondly, minimize the running time of a system because time taken to process the data is a significant one. Normally it takes more time when the number of attributes of a data set is large and sometimes this dataset not supported by all the clustering techniques hence the number of attributes are directly proportional to processing time.

In this paper, PCA is used to reduce the dimension of the data. This is achieved by transforming to a new set of variables (Principal Components) which are uncorrelated and, which are ordered so that the first few retain the most of the variant present in all of the original variables. The first Principal Component is selected to find the initial centroid for the clustering process.

4**. EXISTING METHODS There is no commonly accepted or standard "best" way to determine either the** no. of clusters or the initial starting point values. The resulting set of clusters, both their number and their centroids, depends on the specified choice of initial starting point values. Two simple approaches to cluster initialization are either to select the initial values randomly or to choose the first k samples of the data points.

the new approach was proposed to find the initial centroid using PCA and we compared the results with existing methods. In [19], we have used this method for iris dataset and we have compared the results with other initialization method. This new method was outperformed with better accuracy and less running time than the existing methods. In this paper, we have applied our proposed method for wine, glass and image segmentation dataset. To improve the efficiency of our method we have used heuristics approach to reduce the number of distance calculation in the standard k-means algorithm

5. PROPOSED METHOD The proposed method that performs data partitioning with Principal component. It partitions the given data set into k sets. The median of each set can be used as good initial cluster centers and then assign each data points to its nearest cluster centroid.

steps of the proposed algorithm.

Algorithm 1: The proposed method Steps:

1.Reduce the dimension of the data into d dimension and determine the initial centroid of the clusters by using Algorithm 2

. 2.Assign each data point to the appropriate clusters by using Algorithm

3. In the above said algorithm the data dimensions are reduced and the initial centroids are determined systematically

so as to produce clusters with better accuracy.

Algorithm 2: Dimension reduction and finding the initial centroid using PCA.

Steps: 1.Reduce the D dimension of the N data using Principal Component Analysis (PCA) and prepare another N data with d dimensions (d<=i<=n) to all the centroids cj(1<=j<=k) using Euclidean distance formula..

2. For each data object xi, find the closest centroid cj and assign xi to the cluster with nearest centroid cj and store them in array Cluster[ ] and the Dist[ ] separately. Set Cluster[i] = j, j is the label of nearest cluster. Set Dist[i]= d(xi, cj), d(xi, cj) is the nearest Euclidean distance to the closest center.

3. For each cluster j (1<=j<=k), recalculate the centroids;

4. Repeat 5. for each data-point 5.1 Compute its distance from the centroid of the present nearest cluster

5.2 If this distance is less than or equal to the previous nearest distance, the data-point stays in the cluster Else For every centroid cj Compute the distance of each data object to all the centre Assign the data-point xi to the cluster with nearest centroid cj

6. For each cluster j (1<=j<=k), recalculate the centroids; Until the convergence criteria is met. This algorithm requires two data structure Cluster [ ] and Dist[ ] to keep the some information in each iteration which is used in the next iteration. Array cluster [ ] is used for keep the label if the closest centre while data structure Dist [ ] stores the Euclidean distance of data object to the closest centre.

The information in data structure allows this function to reduce the number of distance calculation required to assign each data object to the nearest cluster, and this method makes the improved k-means algorithm faster than the standard k-means algorithm.

CONCLUSION The main objective of applying PCA on original data before clustering is to obtain accurate results. But the clustering results depend on the initialization of centroid

By using only ICA BSS and UFL using RICA and SFT, clustering accuracy that is better or on par with many deep learning-based clustering algorithms was achieved. For instance, by applying ICA BSS to spectral clustering on the MNIST dataset, we obtained an accuracy of 0.882. This is better than the well-known Deep Embedded Clustering algorithm that had obtained an accuracy of 0.818 using stacked denoising autoencoders in its model.