Empirical Analysis on the Efficiency of Clustering Algorithms Based on the Significance of Cluster Size

Sunitha Cheriyan, Shaniba Ibrahim, Susan Treesa
Higher College of Technology
Muscat, Sultanate of Oman
sunitha.cheriyan@hct.edu.om
shaniba.ibrahim@hct.edu.om
susan@hct.edu.om

ABSTRACT: *This paper mainly focuses on the performance of the various clustering algorithm on a particular dataset based on the number of clusters defined. The analysis is performed on the iris dataset from the dataset library. It also compares the performance of the algorithms based on the number of clusters defined. The various algorithms used for the comparison includes K-Means, Hierarchical, Model based and Density based Clustering based on Statistical models.*

Keywords: Clustering, K-Means, Hierarchical, Model Based, Density Based, efficiency

Received: 14 January 2022, Revised 9 February 2022, Accepted 26 February 2022

DOI: 10.6025/dspaial/2022/1/2/62-72

1. Introduction

Clustering is an important task in data mining application, which has an unsupervised learning technique. K-means clustering, hierarchical clustering, model based clustering and density based clustering are different type of clustering methods. Each clustering method has its own advantages, computational background, theoretical support and disadvantages. Hierarchical clustering algorithms treat each document as a singleton cluster at the outset and then successively agglomerate pairs of clusters until all clusters have been merged into a single cluster that contains all document. Density based clustering is a state of the art clustering technique with numerous application and available in many fields. Model based clustering algorithms are usually represented as an optimization process with an iterative model re-estimation and sample re-assignment. In K-means clustering method depends on the selection of the initial centroids. In this approach k-data elements are selected as initial centers and subsequently distances of all data elements are calculated by a using a Euclidean distance formulae. The important stages of clustering are: select the raw data, use a right kind of algorithm and fix the number of clusters. Identification of the number clusters in each dataset is still persisting as a critical issue. Developing an alternate approach to choose number of clusters that makes limited parametric assumption with the support of a distortion theory is highly effective on a wide range of problems.

2. Relatedwork

Clustering is an unsupervised learning process, which can be done by finding similarities between data based on certain

characteristics found in the data. An efficient clustering technique produces high quality clusters with high intra class similarity and low inter class similarity. As per the various research work done by the researchers, there are the clustering algorithms can be classified into hierarchical based, portioned-based, model based and density based. To determine the exact number of clusters in a particular data set is very significant in cluster analysis. It is very hectic to find out the exact number of required for an analysis, which required an in-depth analysis and visualization of strength of these clusters [1].Density based clustering technique discovers clusters of arbitrary shape. It requires only one input parameters and will support the user to identify an appropriate value [2].Traditional K-means technique can be improved by selecting an initial centroid and later assigning data points by calculating the mean distance between two data points [3]. Majority of hierarchical approaches are agglomerative in nature, start with each observation as its own clusters and groups the observations into an increasingly larger groups [4].

One of the most difficult issues in clustering techniques is to find out the exact number of clusters required the data analysis, these issues can be solved by using an analytic hierarchical process for decision making [5]. DBSCAN is a well-known density based clustering algorithm, the computation issue over dense region of this method also can be effectively handled by using Mr. Scan-algorithm [6].Choosing the number of clusters with limited parametric assumptions can be rigorously motivated by using rate distortion theory, which is effective on a wide range of problem [7].

3. Research Methodology

The main purpose of the research is to evaluate the performance of various clustering techniques and the significance of cluster distances/numbers.

3.1. Data Collection and Preparation

The dataset used for the research is the iris dataset from the dataset library. The dataset contains a set of 150 records under 5 attributes - Petal Length, Petal Width, Sepal Length, Sepal width and Class.

3.2. Methodology

The methods used here includes K-Means, Hierarchical, Model based and Density based Clustering.

3.1.1. K-Means Algorithm

One of the commonly used algorithms in data analysis is K-means algorithm, it is initiated by finding k clusters based on independent variables. It performs division of objects into clusters which are similar between them and are dissimilar to the object belonging to another cluster initially select k random centers and assign objects, which are closer to these centers. In second step recalculate the center of each collection and that forms new k centers. Continue these process until reaches at a point where the center of clusters is hardly moving or reached at a threshold of number of iterative point. The important procedures in this algorithm is: select the k data objects randomly from the original data set, then traverse all data object in the data set by assigning them into a cluster with the highest similarity. The similarity calculation standard calculation is measured by the distance between the data object and the centroid [8], which is usually represented as;

$$D(x,y) = \sqrt{\sum}^{m}_{i=1} (xi, yi)^2 \qquad (1)$$

New centroid of the cluster is recalculated. The minimum error squared sum is used define

$$Z= \sqrt{\sum}^{k}_{j=1} \sqrt{\sum}^{Mi}_{i=1} (si, ci) \qquad (2)$$

The above process is continued in an iterative manner until the standard function converges or the distance between the new and the old center points below a certain threshold. The flow chart for the K-Means algorithm is shown in Figure 1.

3.1.2. Hierarchical Clustering

Hierarchical cluster builds a cluster tree (a dendrogram) to represent data, where each group links to two or more successor groups, which is organized as a tree. These types of analysis outputs a hierarchy, a structure that is more informative than the unstructured set of clusters and does not require to pre specify the number of clusters. The root of the tree has a single cluster

containing all observation and the leaves correspond to individual observations.

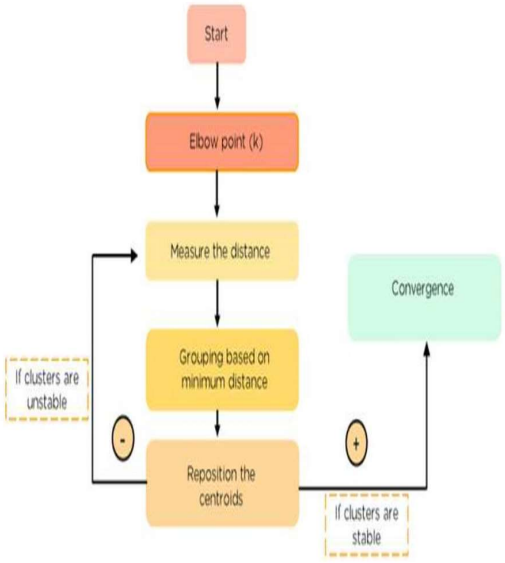


Figure 1. Steps in K-Means Algorithm

Algorithms for hierarchical clustering are usually agglomerative, in which one starts at the leaves and then merges clusters together. It has an added advantage over Kmeans clustering that it results in an attractive tree based observation called dendrogram (a cluster tree) [4]. Hierarchical clustering algorithms are usually found in either agglomerative or divisive types, where agglomerative works on bottom up approach and divisive works on top down approach. It can be performed with either a distance matrix or raw data.

3) Model Based Clustering

The model based clustering assumes that a dataset to be clustered consists of various clusters with different distribution. These types of algorithm use certain models for clusters and tries to optimize the fit between the data and the models, where the data are viewed as coming from a mixture of probability of distribution, each of which represent a different cluster. The clustering assumes a set of n p-dimensional vectors y1, y2……., yn of observations from a multivariate mixture of a finite number of g components or clusters each with some unknown mixing proportions or weights π1, π2…. Πg. The probability density function of finite mixture distributions models can be given by

$$f(yj;\psi) = \Sigma^{g}_{i=1}\ \pi i f\hat{i}\,(yj;\theta i) \qquad (3)$$

Even though model based clustering is a popular tool due to its probabilistic foundation and flexibility, but some cases, it shows disappointing behavior in high-dimensional spaces. Model based clustering assumes a data model and applies an appropriate algorithm to find the most likely model components and the number of clusters. The Figure 2 shows an example of the Model Based Clustering.

4) Density Based Clustering

Density based clustering forms the clusters of densely gathered objects separated by sparse regions and these types of clustering algorithm has a significant role in finding nonlinear shape-structure based on density and working on the basic concepts of density reachability and density connectivity. Density reachability can be explained as: A point p is said to be density reachable from a point q if point p is within e distance from point q and q has sufficient number of points in its neighbors which are within distance 'e'. Density connectivity can be explained as: A point p and q are said to be density connected if there exist a point r which has sufficient number of points in its neighbors and both the point's p and q are within the e distance. This type of clustering forms the clusters of densely gathered objects

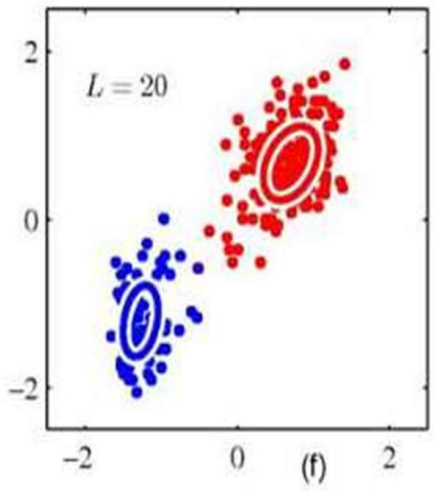
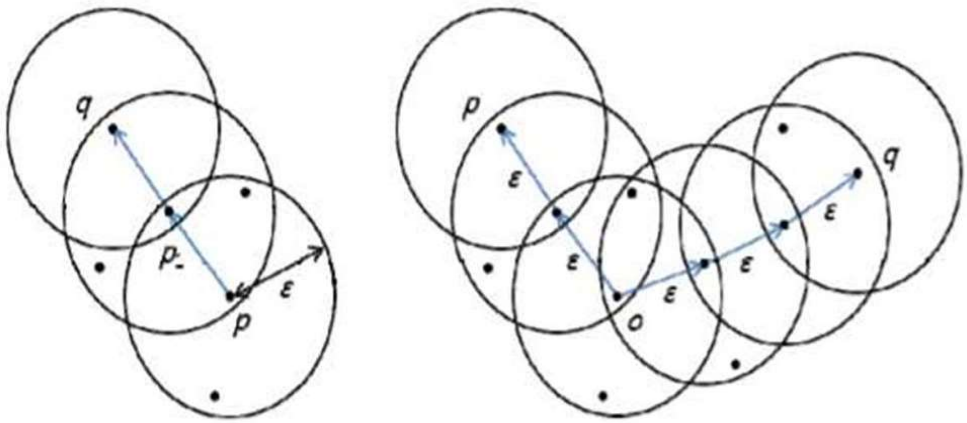


Figure 2 . Model Based Clustering

4) Density Based Clustering

Density based clustering forms the clusters of densely gathered objects separated by sparse regions and these types of clustering algorithm has a significant role in finding nonlinear shape-structure based on density and working on the basic concepts of density reachability and density connectivity. Density reachability can be explained as: A point p is said to be density reachable from a point q if point p is within ε distance from point q and q has sufficient number of points in its neighbors which are within distance 'ε'. Density connectivity can be explained as: A point p and q are said to be density connected if there exist a point r which has sufficient number of points in its neighbors and both the point's p and q are within the e distance. This type of clustering forms the clusters of densely gathered objects separated by sparse region and has the advantage that it can discover the clusters of arbitrary shapes and also filter out noise objects [2], [6].



(a) q is density-reachable from p. (b) p and q are density-connected with each other.

Figure 3 . Density Based Clustering

4. Experimeand Analysis

The set of observations that are similar in patterns are grouped together as clusters. The software used for the analysis is R. The 3D visualization of the iris dataset is as shown in the Figure 4.

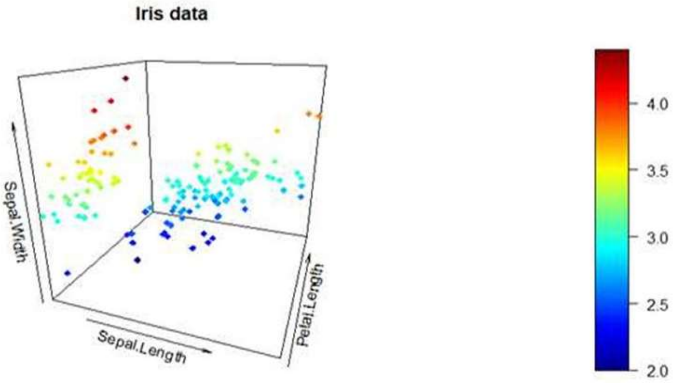


Figure 4. Iris Dataset

4.1. K-Means Clustering

In this model, the iris dataset is loaded and K- Means clustering is applied to the dataset. This Clustering can be used on unlabeled data and is an algorithm of unsupervised machine learning. The correlations of attributes are shown in the table 1 below:

Attributes	Petal.Length	Petal.Width	Sepal.Length	Sepal.Width
Petal.Length	1	0.963	0.872	-0.428
Petal.Width	0.963	1	0.818	-0.366
Sepal.Length	0.872	0.818	1	-0.118
Sepal.Width	-0.428	-0.366	-0.118	1

Table 1. Correlation Table

The correlation table shows that there is a strong association between petal.width and sepal.length.

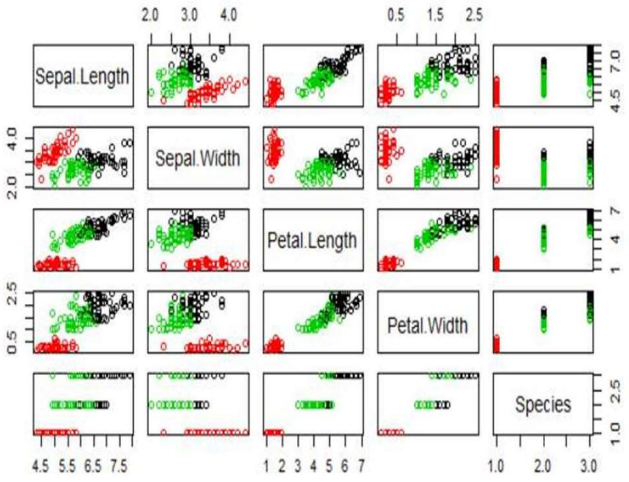


Figure 5. K Means for k = 3

The clustering is performed for k = 3, 4. The Figure 5 shows the distribution of clusters for k=3.

Afterwards, a dimensionality reduction is performed in order to support the visualization of the data in two dimensions.

The number of clusters is 3 and the distance measure is taken by squared Euclidean distance. The average cluster distance is 0.927 and the Davies-Bouldin index. It was found that at Cluster 0 the average distance was about 0.985. The petal.width is on average 71.18% larger, petal.length is on average 63.99% larger, sepal.length is on average 62.43% larger. In Cluster 1 the average distance was about 0.864. The sepal.width is on average 35.99% smaller, petal.length is on average 24.05% larger. The petal.width is on average 21.50% larger. In Cluster 2 petal.width is on average 86.72% smaller. petal.length is on average 83.25% smaller, sepal.length is on average 54.25% smaller. The heat map shown in Figure 6 shows the cluster details.

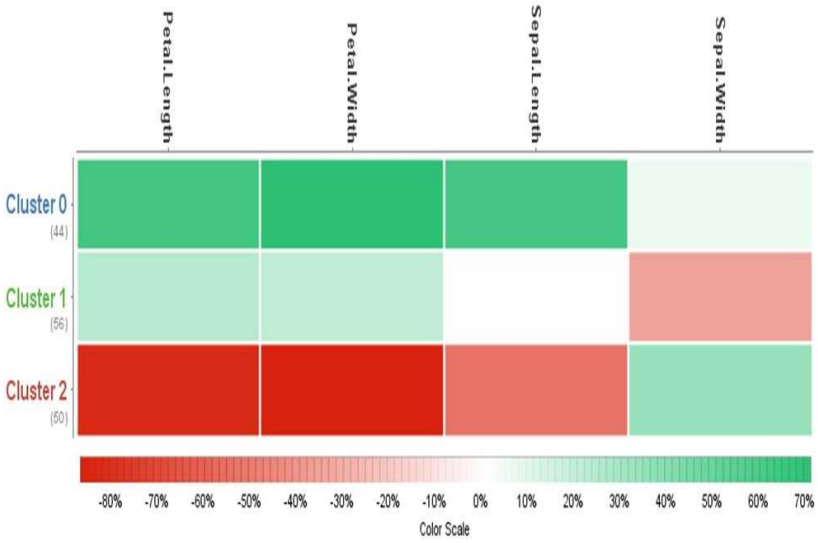


Figure 6. Heat Map

Secondly, the K Means was performed for k=4 and the centroid table is as shown in the table 2. The average Cluster Distance found is 0.758 and the Davies-Bouldin Index is 0.907. In Cluster 0, a total of 50 items were grouped with an average distance of 0.592. The petal.width is on average 37.90% larger, petal.length is on average 37.42% larger, sepal.length is on average 19.09% larger. In Cluster 1, a total of 49 items were grouped with an average distance of 0.819. The petal.width is on average 86.82% smaller, petal.length is on average 83.13% smaller. sepal.length is on average 53.59% smaller. In Cluster 2, petal.width is on average 84.12% larger, sepal.length is on average 74.72% larger, petal.length is on average 74.04%. In Cluster 3, a total of 22 items were grouped with an average distance of 0.775. The sepal.width is on average 58.73% smaller, sepal.length is on average 22.54% smaller, petal.width is on average 3.66% smaller.

Cluster	Petal.Length	Petal.Width	Sepal.Length	Sepal.Width
Cluster 0	0.585	0.547	0.356	-0.393
Cluster 1	-1.299	-1.252	-0.999	0.903
Cluster 2	1.157	1.213	1.393	0.232
Cluster 3	0.039	-0.053	-0.420	-1.425

Table 2. Centroid Table

The Figure 7 shows the performance of the algorithms for the value of k=4. It indicates that there is no significance in grouping as there are overlapping in some groups. For this dataset the performance is better with value of $k = 3$.

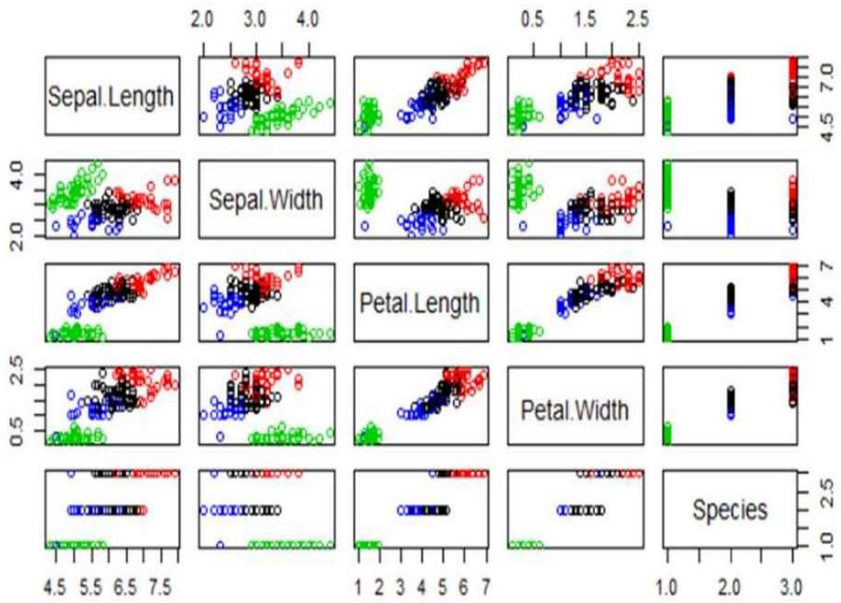


Figure 7. K Means for k = 4

4.2. Hierarchichal Clustering

The Figure 8 displays the outcome of the ward. D2, and we see the same visual result in both cases. That is, the dendrogram and the distribution as scatter plot. The scatter plot for the dataset is shown in Figure 9. The distance is calculated in Euclidean distance.

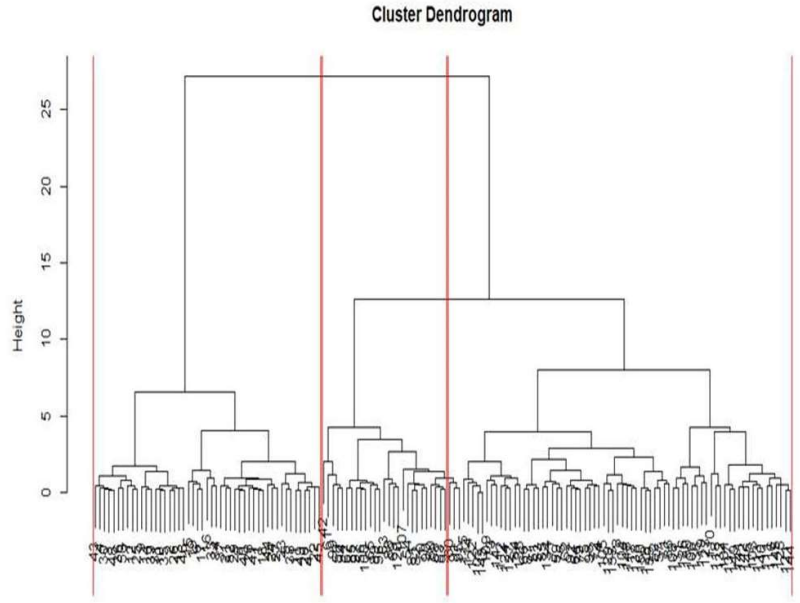


Figure 8. Dendrogram of Hierarchical Clustering

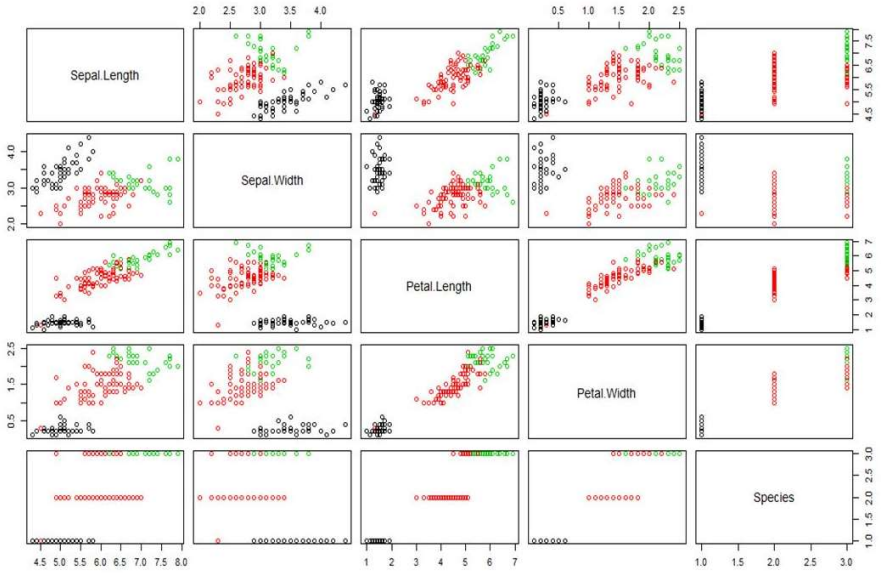


Figure 9. Scatter Plot for Hierarchical Clustering

4.3. Model Based Clustering

The model based clustering can be based on BIC, classification, uncertainty and density. The Figure 10a displays the BIC for parameterized Gaussian mixture models fitted by EM algorithm initialized by model-based clustering. 10b, c, d represents the Classification, uncertainty and density respectively.

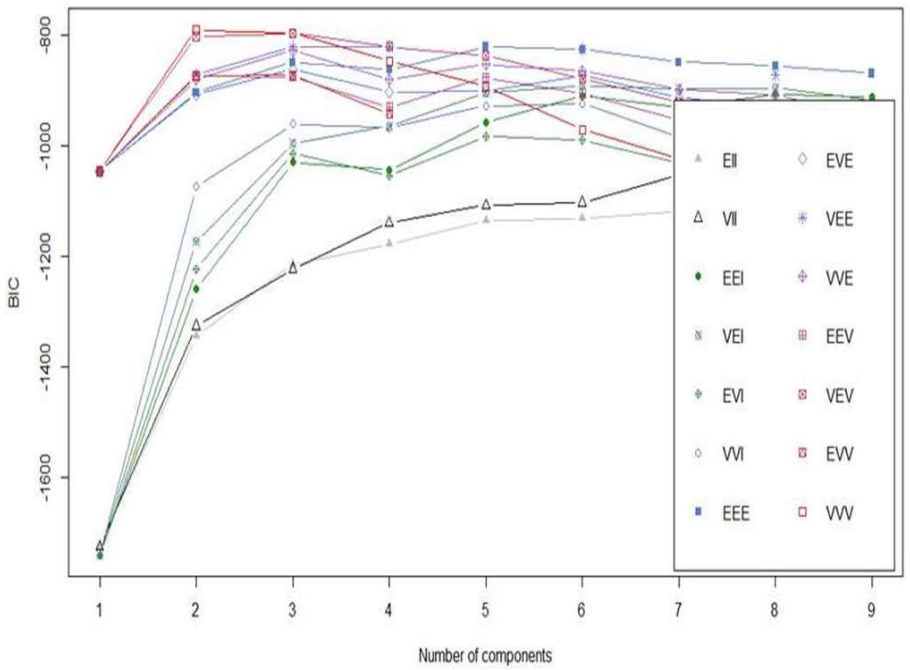


Figure 10a. BIC

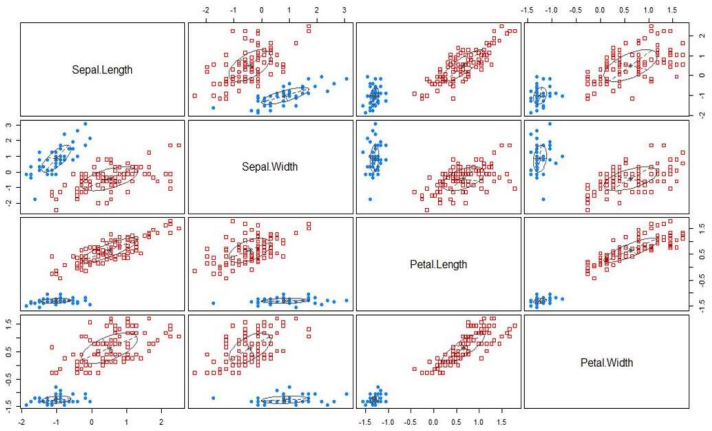


Figure 10b. Classification

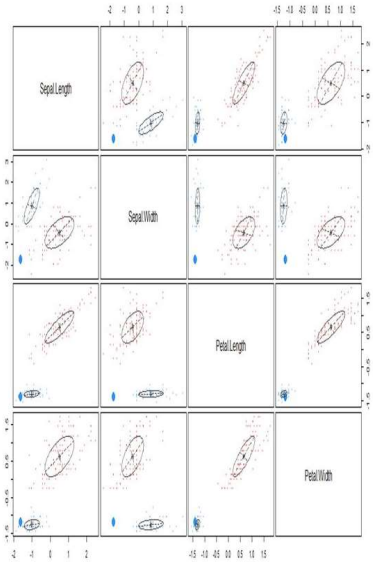


Figure 10c. Uncertainty

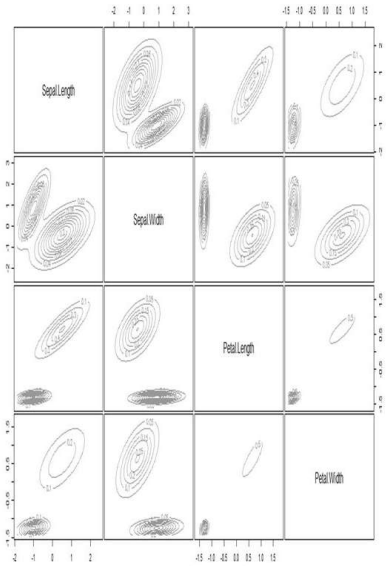


Figure 10d. Density

4.4. Density based Clustering

The density based clustering is performed on the dataset. It was observed that the eps value obtained is 0.7 with minimum points 5. The Figure 11 indicates the model based on DBScan performed on the dataset. It identifies the area were the population is dense. Min points is the dimensions in the data frames. The dataset contains 5 variables so the min point value is set to 5. After performing the test, it was observed that it identified 2 different clusters and 6 noise points. The noise points represent the points far away.

5. Conclusion

In the test, we have carried out four continuous variables. Like Model Based Clustering the density based clustering also separates clusters with two clusters. The numeric variables are similar to each other and are not completely separated. We need to get more observations so that we can sample more data. So the population can be easily separated. As the dimensions in the dataset is increased, the need of numbers of observations also increases exponentially. So the PCO analysis can be used to identify the clusters and adequately separate the clusters accurately. So the different algorithms performs well in different datasets. K-Means clustering is scalable with large datasets, the cluster shape is hyper spherical and works well with numeric

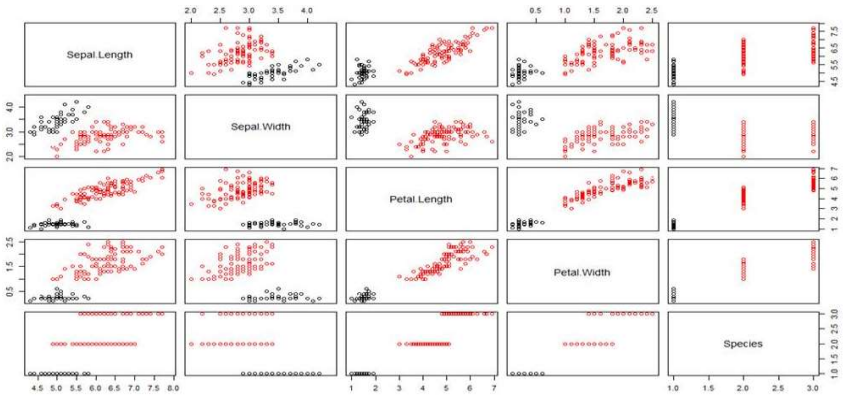


Figure 11 . Density Based Clustering

data. It is sensitive to noise and outliers. The efficiency is O(n,k,t) where n,k,t are the number of iterations, clusters and data points respectively. But hierarchical clustering is not scalable with large datasets. The efficiency is represented in the notation O(n2). Whereas Model based clustering can handle large databases and the cluster shape is ellipsoid when mixture model is applied. Density based clustering does not work with high dimensional data, the cluster shape is arbitrary and handles noise. The complexity is denoted as O(nlogn).

References

[1] Benson-Putnins, D. A. V. I. D., Bonfardin, M., Magnoni, M. E., & Martin, D. (2011). Spectral clustering and visualization: a novel clustering of fisher's iris data set. *SIAM Undergraduate Research Online*, 4.

[2] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. *In Kdd* (Vol. 96, No. 34, pp. 226-231).

[3] Kumar, S., & Asger, M. Analysis Clustering Techniques in Biological Data with R.

[4] Sembiring, R. W., Zain, J. M., & Embong, A. (2011). A comparative agglomerative hierarchical clustering method to cluster implemented course. arXiv preprint arXiv:1101.4270

[5] Akogul, S., & Erisoglu, M. (2017). An Approach for Determining the Number of Clusters in a Model-Based Cluster Analysis. *Entropy*, 19(9), 452.

[6] Welton, B., Samanas, E., & Miller, B. P. (2013, November). Mr. scan: Extreme scale density-based clustering using a tree-based network of gpgpu nodes. *In High Performance Computing, Networking, Storage and Analysis (SC)*, 2013 International Conference for (pp. 1-11). IEEE

[7] Zhong, S., & Ghosh, J. (2003, May). Scalable, balanced model-based clustering. In Proceedings of the 2003 *SIAM International Conference on Data Mining (pp. 71-82)*. Society for Industrial and Applied Mathematics.

[8] Huang, Q., & Zhou, F. (2017, March). Research on retailer data clustering algorithm based on spark. *In AIP Conference Proceedings* (Vol. 1820, No. 1, p. 080022). AIP Publishing.

[9] Purkait, G., & Singh, D. (2017). An effort to optimize the error using statistical and soft computing methodologies. *Journal of Applied Computer Science & Artificial Intelligence*, 1(1), 15-20.

[10] Soni, K. G., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *International Journal of Computational Intelligence Research*, 13(5), 899-906.

[11] Gan, J., & Tao, Y. (2017, May). Dynamic Density Based Clustering. In Proceedings of the 2017 *ACM International Conference on Management of Data* (pp. 1493-1507). ACM.

[12] based clustering using a tree-based network of gpgpu nodes. In High Performance Computing, *Networking, Storage and Analysis (SC), 2013 International Conference for (pp. 1-11)*. IEEE.

[13] Zhang, H., Thieling, T., Prins, S. C. B., Smith, E. P., & Hudy, M. (2008). Model-Based Clustering in a Brook Trout Classification Study within the Eastern United States. Transactions of the American Fisheries Society, 137(3), 841-851.

[14] Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, 98(463), 750-763.

[15] Nagpal, P. B., & Mann, P. A. (2011). Comparative study of density based clustering algorithms. *International Journal of Computer Applications*, 27(11), 421-435.

[16] Mai, S. T., Assent, I., & Storgaard, M. (2016, August). AnyDBC: an efficient anytime density-based clustering algorithm for very large complex datasets. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1025-1034). ACM.

[17] Chen, C. C., & Chen, M. S. (2015). HiClus: Highly Scalable Densitybased Clustering with Heterogeneous Cloud. *Procedia Computer Science*, 53, 149-157.