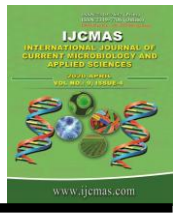
Int.J.Curr.Microbiol.App.Sci (2020) *9*(4): 2952-2961

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 9 Number 4 (2020)**
Journal homepage: http://www.ijcmas.com



Original Research Article https://doi.org/10.20546/ijcmas.2020.904.346

Clustering and Validation of *Iris* Flower Dataset using Relative Criteria

K. Sujatha[1]*, Vijayakumar Selvaraj[2] and N. Nevashini[3]

[1]Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia, West Bengal, India
[2]Manager Learning Analytics, Imarticus Learning Chennai, Chennai, Tamil Nadu, India
[3]Technology Lead-Data Science, iNurture Education Solutions Private Limited, Bengaluru, Karnataka, India

**Corresponding author*

Keywords

Iris flower, Clustering, Dunn, DB, RMSSTD and RS index

Article Info

Accepted:
22 March 2020
Available Online:
10 April 2020

A B S T R A C T

In the present work, attempts have been made to analyze the *Iris* data set with clustering technique which is the 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. The *Iris* flower data set is a popular multivariate data set introduced by Sir Ronald Fisher as an example of discriminant analysis. The data on four characteristics of the three species of the *Iris* Flower, sepal length, sepal width, petal length and petal width has been taken from https://ieeexplore.ieee.org/document/771092 and has been analyzed using SAS software. Here we have extended the algorithm for better visualization of possible cluster structures and also to validate clusters. The optimal number of clusters was found in this dataset by using the four cluster validity indices viz., Dunn, DB, RMSSTD and RS indices which yield three and this is configurable to the real partitions of the dataset.

Introduction

The *Iris* Flower Dataset is a popular multivariate dataset that was introduced by R.A. Fisher as an example for discriminant analysis. The data reports on four characteristics of the three species of the *Iris* Flower, sepal length, sepal width, petal length and petal width. There are many different statistical techniques available for clustering i.e.grouping data into categories based on some measure of inherent similarity or distance (unsupervised learning) and classification which is a technique where predefined labels are assigned to instances by properties (supervised learning) (Fraix Burnet

et al., 2015; De *et al*., 2013). Cluster analysis organizes data into groups based on similarities between the data points. Sometimes the data contains natural divisions that indicate the appropriate number of clusters. Other times, the data does not contain natural divisions, or the natural divisions are unknown. In such a case, one might determine the optimal number of clusters to group the data. To determine how well the data fits into a particular number of clusters, compute index values using different evaluation criteria, such as Dunn, DB, RMSSTD and RS index. Visualize clusters by creating a dendrogram plot to display a hierarchical binary cluster tree. In the present work, attempts have been made to explore *Iris*data for clusters using hierarchical clustering scheme and validation of clustering using relative criteria.

Materials and Methods

The *Iris* flower data on four characteristics of the three species of the *Iris* Flower, sepal length, sepal width, petal length and petal width has been taken from https://ieeexplore.ieee.org/document/771092 and has been analyzed using SAS software. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects clustered. There are two major methods of clustering – hierarchical and k-means clustering. Hierarchical Clustering groups data over a variety of scales by creating a cluster tree or *dendrogram*. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows you to decide the level or scale of clustering that is most appropriate for the application. K-Means Clustering is a partitioning method. The function k-means partitions data into *k* mutually exclusive clusters and returns the index of the cluster to which it has assigned each observation.

By and large, the whole process of Hierarchical Clustering accomplished in two phases.

Phase 1: Working out the distance

Working out the distance between any two individuals considering all the characters under study together.

Distance function

If there are two objects (P,Q) with their observations X and Y, then d(P,Q) is a distance function if it has the following properties: (i) Symmetry: d(P,Q) = d(Q,P); (ii) Non-Negativity: d(P,Q)≥0; (iii) Definiteness: d(P,Q)=0 if &only if P=Q; (iv) Triangle inequality: d(P,Q)≤d(P,R)+d(R,Q). Some of the commonly applied distance measures are Euclidean Distance, Squared Euclidean Distance, Minkowski's Metric, Mahalanobis distance. Clusters are formed with the help of Euclidean distance measures because it is easy to compute and understand.

Euclidean Distance: D(x,y) =

$$\sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

where X and Y are the units measured over i=1,2, …,m characters.

Phase 2: Linking

Joining the similar elements into one group and relatively dissimilar elements into different groups.

Cluster process

Assign each object to a separate cluster.
Evaluate all pair-wise distances between

clusters.
Construct a distance matrix using the distance values.
Look for the pair of clusters with the shortest distance.
Remove the pair from the matrix and merge them.
Evaluate all distances from this new cluster to all other clusters, and update the matrix.
Repeat until the distance matrix is reduced to a single element.

Cluster validation

Clustering validation, which evaluates the goodness of clustering results, has long been recognized as one of the vital issues essential to the success of clustering applications. It has four main components:

Determine whether there is a non-random structure in the data;
Determine the number of clusters;
Evaluate how well a clustering solution fits the given data when the data is the only information available;
Evaluate how well a clustering solution agrees with partitions obtained based on other data sources.

Measure of cluster validity

There are three different techniques for evaluating the result of the clustering algorithms (Theodoridis and Koutroubas, 1999):

External criteria

Used to measure the extent to which cluster labels match externally supplied class labels. Example: Entropy

Internal criteria

Used to measure the goodness of a clustering structure without respect to external information. Example: Sum of Squared Error (SSE)

Relative criteria

The basis of the relative criteria is the comparison of the different clustering schema. One or more clustering algorithms are executed multiple times with different input parameters on the same data set. Hence the relative criteria are considered to be better than both external and internal criteria. The relative criteria aim is to choose the best clustering schema from different results.

Two measurement criteria have been proposed for evaluating and selecting an optimal clustering scheme (Berry and Linoff, 1996):

Compactness: The member of each cluster should be as close to each other as possible. A common measure of compactness is the variance.

Separation: The clusters themselves should be widely separated.

Cluster validity indices

Dunn index

Dunn index looks to identify compact and well-separated clusters (Dunn. J., 1974) and is defined as

$$D_c = \min_{i=1\ldots n}\left\{\min_{j=i+1\ldots n}\left(\frac{d(C_i, C_j)}{\max_{k=1\ldots n}(diam(C_k))}\right)\right\}$$

where, $d(C_i,C_j)$ is the inter-cluster distances between clusters C_i and $C_j$and $diam(C_j)$ is the diameter of an i^{th} cluster.

If the dataset contains compact and well-separated clusters, the distance between the

clusters (inter-cluster distance) is expected to be large and the diameter of the clusters (intracluster distance) is expected to be small. Thus, based on Dunn's index definition, the large value of the index indicate the presence of compact and well-separated clusters. The index D_c does not exhibit any trend concerning C; hence the maximum in the plot of D_c versus C can be used to indicate the number of clusters. Dunn index for a single cluster is zero.

Davies-Bouldin (DB) Index

DB index is defined in the following equation (D.L. Davies and D.W. Bouldin, 1979)

$$DB_n = \frac{1}{n}\sum_{i=1}^{n} R_{ij}$$

where n is the number of clusters in the data set and R_{ij} is the average of similarity measure between the clusters Ci and Cj and can be defined as R_{ij} = max of $\left(\frac{diam(C_i)+diam(C_j)}{d(C_i,C_j)}\right)$ where $d(C_i,C_j)$ is the inter-cluster distance and *diam (Ci)* is the diameter of the ith cluster.

DB index measures the similarity between two clusters and it should be less. If the DB index exhibits no trends to the number of clusters, then the minimum value of DB in its plot versus the number of clusters will be considered the optimum number of clusters. Similarity index R_{ij} between C_i and C_j is defined to satisfy the following condition:

1. $R_{ij} \geq 0$
2. $R_{ij} = R_{ij.}$
3. S_i is the measure of dispersion of a cluster $C_{i.}$;if $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$.
4. If $s_i > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik.}$
5. If $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik.}$

These condition state that the R_{ij} is nonnegative and symmetric if both clusters, C_i and $C_{j,}$ collapse to a single point, then $R_{ij} = 0$.

RMSSTD (Root Mean Square Standard Deviation) and RS (RSquared) indices

RMSSTD was proposed by Subhash Sharma in 1996. It is the measure of homogeneity within clusters. Lower the RMSSTD value indicates the higher the homogeneity.

RS indicates the extent to which clusters are different from each other. RS values range from0 to 1. If the value is 0, then this indicates that there are no differences among the clusters and1 indicates that there are significant differences among the clusters. RMSSTD can be defined as

$$\text{RMSSTD} = \sqrt{\frac{\sum_{i=1,\ldots n}\sum_{\substack{j=1,\ldots d}}^{n_{ij}}(x_k-\bar{x})^2}{\sum_{i=1,\ldots n}\sum_{\substack{j=1,\ldots d}}(n_{ij}-1)}}$$

where*n* is the number of clusters.

RMSSTD index can further be simplified in the following equation

$$\text{RMSSTD} = \sqrt{\frac{SS_1+SS_2+\cdots SS_p}{df_1+df_2+\cdots df_p}}$$

where, SSj , (j=1,2,…,p) is within sum of square(SS_{within}) of j^{th} attribute of the cluster objects. And this, will be calculated by using the following equation

$$SS_j = \sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2$$

RS index will be calculated by taking the ratio of $SS_{between}$ and SS_{total}.

$$RS = \frac{SS_{between}}{SS_{total}} = \frac{SS_{total}-SS_{within}}{SS_{total}}$$

Where SS_{total} is defined in given equation

$$SS_{total} = \sum_{j=1}^{p} \sum_{j=1}^{n} (X_{ij} - \bar{X}_j)^2$$

Calculate index value in each stage of clustering algorithm and plot the values as against the number of clusters. If knee point occurs in the graph then the lowest value, which makes knee point, of the index that corresponds to the cluster number considered as the optimum number of cluster. If there is no trend occurs as against the number of cluster then the minimum value of the index will be taken.

Results and Discussion

Clustering techniques

Clustering processwas easily carried out by Euclidean distance between the cases using agglomerative method of hierarchical clustering. Initially, every single data point is considered as the single cluster and a further cluster is formed between the cases using single linkage i.e. nearest neighbor. From the analysis, the first cluster was formed between the cases 143 and 102 at a distance of 0.0 followed by cases 1 and 18 of distance 0.05 consisting of two members.

Consequently, the whole *Iris*data is formed into a single cluster by joining the case34 consisting of one member with the case 23 consisting of 119 members. It is delineated that, the distance between the clusters is gradually increased for succeeding clusters. These are represented in the dendrogram tree which is illustrated in figure 1.

Cluster validity indices

Cluster validation is necessary for finding the optimum number of clusters in the given data set and also to find whether the results obtained from the clustering algorithm may or may not be configurable to the real data set. Validation of cluster is done by Cluster validity indices namely Dunn index, Davies Bouldin index, RMSSTD and RS indices.

Dunn index

Dunn index value is negligible till 136th cluster which is formed between the cases 119 and 131 consisting of 87 members and for the last cluster i.e., 149th cluster's Dunn index value is negligible. Hence it is not included in the table. The Dunn index value for 137th cluster which is formed between the cases 119 and 135 is 0.130 consisting of 88 members at the distance of 0.269. The value of the last cluster i.e., 148th cluster is formed between the cases 107 and 23 is 0.223. From table 1, it is observed that the largest Dunn index value is 0.233 for the 147th cluster formed between the cases 107 and 93. A large value of the index indicates the presence of compact and well-separated clusters (Fig. 2).

Davies Bouldin index

The DB index exhibits no trends for the number of clusters and thus we seek the minimum value of DB in its plot versus the number of clusters (Halkidi *et al.*, 2001). The DB index value is negligible till 136th cluster which is formed between the cases 119and 131 consisting of 87 members and for the last cluster i.e., 149th cluster's DB index value is negligible. The DB index value for 137th cluster which was formed between the cases 119 and 135 is 0.558 consisting of 88members at the distance of 0.269. The value of the last cluster i.e., 148th cluster is formed between the cases 107 and 23 is1.327. From table 2, it is observed that the smaller DB index value is 0.515for the 147th cluster formed between the cases 107 and 93.The smaller value of Davies Bouldin index indicates better cluster configuration. DB index value for the given cluster is illustrated in table 3 and figure 3.

Table.1 Dunn index value table

Cluster joining		At distance	No. of members	Dunn index
Case 119	Case 135	0.269	88	0.130
Case 136	Case 119	0.269	89	0.132
Case 23	Case 16	0.274	48	0.134
Case 93	Case 56	0.278	2	0.134
Case 136	Case 109	0.278	90	0.150
Case 23	Case 42	0.312	49	0.152
Case 110	Case 136	0.316	91	0.156
Case 110	Case 61	0.324	95	0.152
Case 107	Case 110	0.367	96	0.169
Case 107	Case 132	0.409	98	0.178
Case 107	Case 93	0.430	100	0.233
Case 23	Case 107	0.624	149	0.223

Table.2 DB index value table

Cluster joining		At distance	No. of members	DB index
Case 119	Case 135	0.269	88	0.558
Case 136	Case 119	0.269	89	0.568
Case 23	Case 16	0.274	48	0.585
Case 93	Case 56	0.278	2	0.623
Case 136	Case 109	0.278	90	0.543
Case 23	Case 42	0.312	49	0.563
Case 110	Case 136	0.316	91	0.563
Case 110	Case 61	0.324	95	0.604
Case 107	Case 110	0.367	96	0.520
Case 107	Case 132	0.409	98	0.519
Case 107	Case 93	0.430	100	0.515
Case 23	Case 107	0.624	149	1.327

Table.3 RMSSTD index value table

Cluster joining		At distance	No. of members	RMSSTD index
Case 119	Case 135	0.269	88	0.521
Case 136	Case 119	0.269	89	0.528
Case 23	Case 16	0.274	48	0.260
Case 93	Case 56	0.278	2	0.197
Case 136	Case 109	0.278	90	0.528
Case 23	Case 42	0.312	49	0.273
Case 110	Case 136	0.316	91	0.533
Case 110	Case 61	0.324	95	0.572
Case 107	Case 110	0.367	96	0.574
Case 107	Case 132	0.409	98	0.596
Case 107	Case 93	0.430	100	0.621
Case 23	Case 107	0.624	149	1.070
Case 34	Case 23	0.789	150	1.068

Table.4 RS index value table

Clusters Joining		No. of Members	RS index
Case 136	Case 109	90	0.889554
Case 23	Case 42	49	0.878838
Case 110	Case 136	91	0.876102
Case 110	Case 61	95	0.872748
Case 107	Case 110	96	0.848114
Case 107	Case 132	98	0.844973
Case 107	Case 93	100	0.761589
Case 23	Case 107	149	0.744051
Case 34	Case 23	150	0.731269

Scale: X axis 1cm = 5.0 units
Y axis 1 cm = 0.2 unit

Fig.1 Dendrogram for *Iris* flower data set

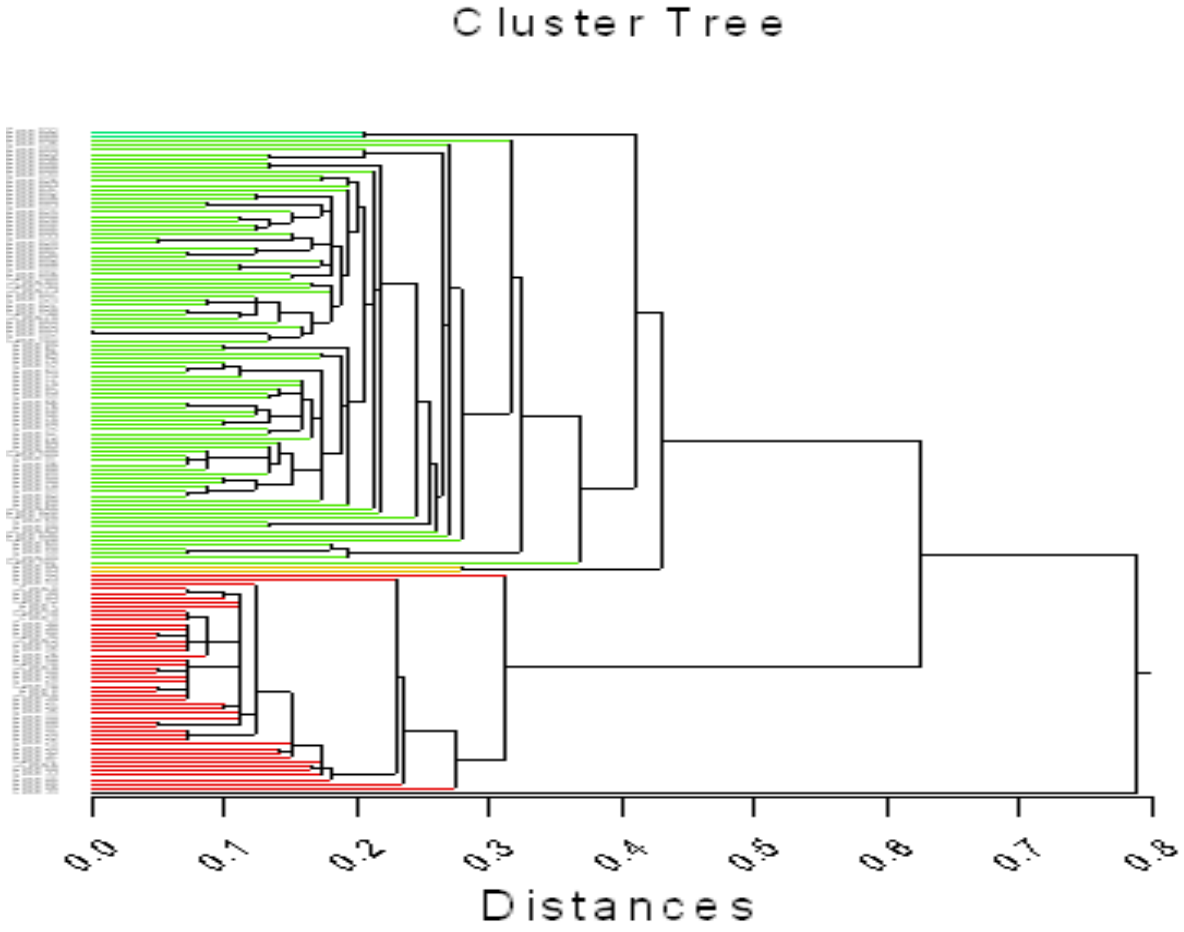
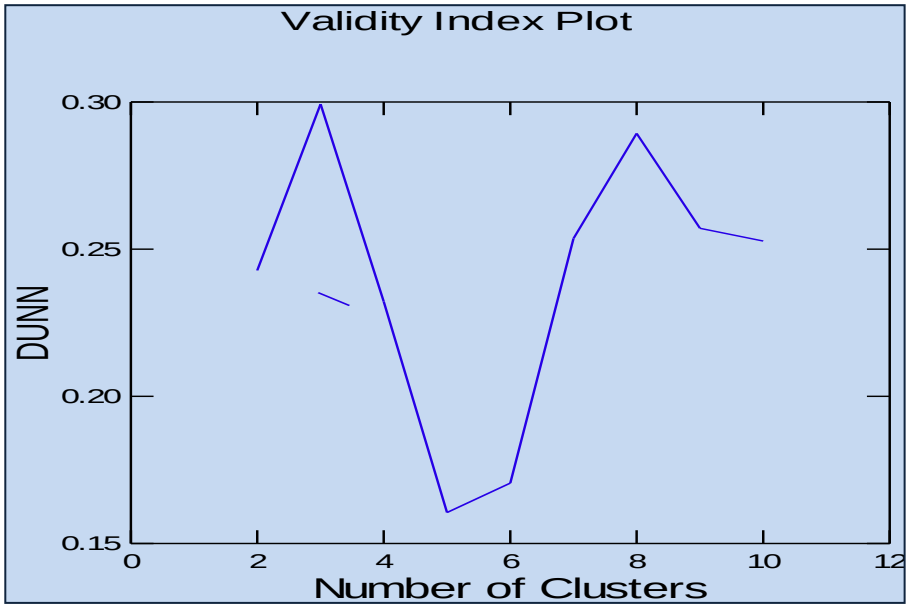


Fig.2 DunnIndex graph

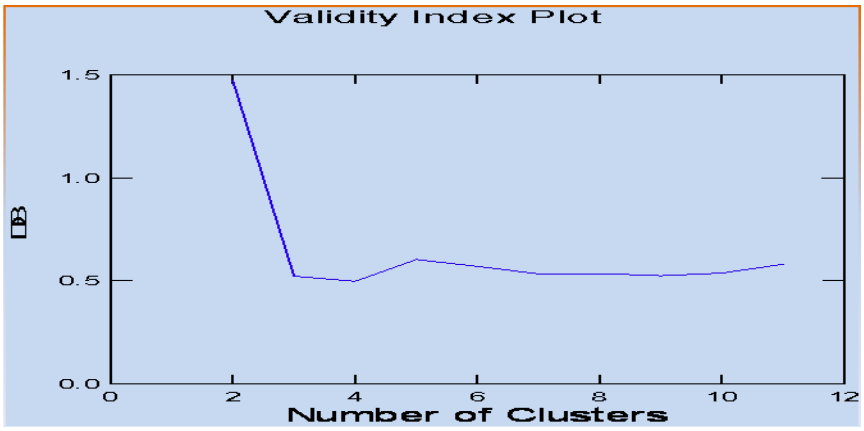


Scale:

X axis 1cm = 2.0 units

Y axis 1 cm = 0.05 unit

Fig.3 DB index graph

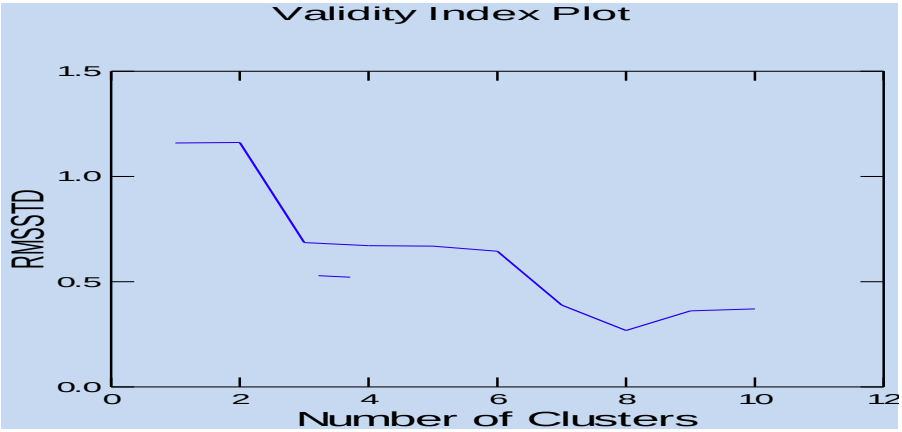


Scale:

X axis 1cm = 2.0 units

Y axis 1 cm = 0.5 unit

Fig.4 RMSSTD Index Graph

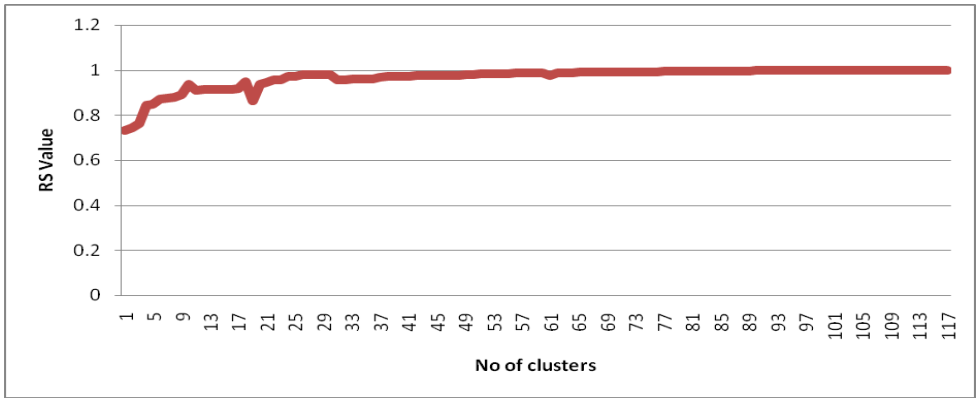


Scale:

X axis 1cm = 2.0 units

Y axis 1 cm = 0.5 unit

Fig.5 RS index graph



RMSSTD index

RMSSTD index value is negligible till 136th cluster which is formed between the cases 119 and 131 consisting of 87 members. The RMSSTD index value for 137th cluster which is formed between the cases 119 and 135 is 0.521 consisting of 88members at the distance of 0.269. The value of the last cluster i.e., 149thcluster is formed between the cases 23 and

34 is1.068consisting of 150 members at the distance of 0.789. From table 3, it is apparent that the smallest RMSSTD index value is 0.197for the 140th cluster formed between the cases 93 and 56. Lower the RMSSTD value indicates the higher homogeneity (Fig. 4).

RS index

Unlike other indices, the value of RS index value is not negligible as the value of RS index gradually increases as shown in graph 5**.** The RS index value for 149th cluster (i.e. single cluster) which is formed between the cases 34 and 23 is 0.731269consisting of 150 members. The value of the succeeding cluster increases gradually. But there is a sudden increase in 145th cluster (i.e. four clusters) formed between the cases 107 and 132, whose value is0.844973. This sudden increase in the value of RS index leads to the formation of highest knee point in the 147th cluster (i.e. three clusters). RS index indicates the extent to which clusters are differing from each other. Higher the RS index value indicates the better cluster configuration in the data set (Fig. 5 and Table 4).

In conclusion, this work has found a way to exploit the true use of clustering by applying clustering algorithms on unclassified datasets to generate clusters for them as well as validating the clusters by the application of indexing techniques.For the given *Iris* flower data set the optimum number of cluster was found out to be three. Cluster Validity Indices viz., Dunn, DB, RMSSTD and RS validity indices showed that the results are close to the real partitions of the data set. By using this pattern and classification, the unknown data can be classified more precisely in many fields.

References

Berry, M.J.A., and Linoff, G. 1996. Data mining techniques for marketing, sales and customer support. John Wiley & Sons, Inc., USA.

Davies, D.L., and Bouldin, D.W. 1979. "A cluster separation measure", IEEE Transactions on pattern Recognition and machine intelligence, 1(2), 224-227.

De, T., Chattopadhyay, T., and Chattopadhyay, A. K. 2013. Comparison among clustering and classification techniques on the basis of galaxy data. Calcutta Stat. Assoc. Bull. 65, 257-260.

Dunn, J.C., 1974. Well Separated Clusters and Optimal Fuzzy Partitions, Journal of Cybernetics,4, 95-104.

Fraix-Burnet, D., Thuillard, M, and Chattopadhyay, A. K. 2015. Multivariate approaches to classification in extragalactic astronomy, Frontiers in Atronomy and Space Science, 2, 1-17.

Halkidi, M., Vazirgiannis, M., and Batistakis. I. 2000. Quality Scheme Assessment in the Clustering Process. In proceedings of PKDD, Lyon, France.

https://ieeexplore.ieee.org/document/771092.

James C. Bezdek, James M. Keller, Raghu Krishnapuram, Ludmila I. Kuncheva, and Nikhil R. Pal. 1999. Will the *Real* Iris Data Please Stand Up? IEEE Transactions on Fuzzy Systems 7(3): 368-369.

Theodoridis, S, and Koutroubas, K. 1999. Pattern Recognition, Academic Press.

How to cite this article:

Sujatha, K., Vijayakumar Selvaraj and Nevashini, N. 2020. Clustering and Validation of *Iris* Flower Dataset using Relative Criteria. *Int.J.Curr.Microbiol.App.Sci.* 9(04): 2952-2961.
doi: https://doi.org/10.20546/ijcmas.2020.904.346