Performance Analysis of Clustering Algorithm in Data Mining in R Language

Avulapalli Jayaram Reddy[1(✉)], Balakrushna Tripathy[2], Seema Nimje[1], Gopalam Sree Ganga[1], and Kamireddy Varnasree[1]

[1] School of Information Technology and Engineering, VIT, Vellore, India
ajayaramreddy@vit.ac.in, seemadnimje@gmail.com
[2] School of Computer Science and Engineering, VIT, Vellore, India

Abstract. Data mining is the extraction of different data of intriguing as such (constructive, relevant, constructive, previously unexplored and considerably valuable) patterns or information from very large stack of data or different dataset. In other words, it is the experimental exploration of associations, links, and mainly the overall patterns that prevails in large datasets but is hidden or unknown. So, to explore the performance analysis using different clustering techniques we used R Language. This R language is a tool, which allows the user to analyse the data from various and different perspective and angles, in order to get a proper experimental results and in order to derive a meaningful relationships. In this paper, we are studying, analysing and comparing various algorithms and their techniques used for cluster analysis using R language. Our aim in this paper, is to present the comparison of 5 different clustering algorithms and validating those algorithms in terms of internal and external validation such as Silhouette plot, dunn index, Connectivity and much more. Finally as per the basics of the results that obtained we analyzed and compared, validated the efficiency of many different algorithms with respect to one another.

1 Introduction

R utilizes accumulations of bundles to perform diverse capacities. CRAN venture sees give various bundles to various clients as per their taste. R bundle contain diverse capacities for information mining approaches. This paper looks at different bunching calculations on Hepatitis dataset utilizing R. These grouping calculations give diverse outcome as indicated by the conditions. Some grouping methods are better for huge informational index and a few gives great come about for discovering bunch with subjective shapes. This paper is wanted to learn and relates different information mining grouping calculations. Calculations which are under investigation as takes after: K-Means calculation, K-Medoids, Hierarchical grouping algorithm, Fuzzy bunching and cross breed bunching. This paper contrasted all these grouping calculations agreeing with the many elements. After examination of these grouping calculations we depict what bunching calculations ought to be utilized as a part of various conditions for getting the best outcome.

I. Zelinka et al. (Eds.): ICSCS 2018, CCIS 837, pp. 364–372, 2018.
https://doi.org/10.1007/978-981-13-1936-5_39

2 Related Work

Few of the researches have worked on different algorithms and implemented few of them, as per that while others have worked on the existing algorithm few have implemented the new one's. applied various indices to determine the performance of various clustering techniques and validating the clustering algorithms.

3 Clustering Analysis Using R Language

Data mining is not performed exclusively by the application of expensive tools and software, here, we have used R language. R is a language and it's a platform for statistical computing and graphics. The clustering techniques which we used here are of basically four types, Partitioning methods, Hierarchal methods, Model based methods, Hybrid Clustering. Here hepatitis dataset is used to validate the results.

4 Clustering Concepts

Clustering analysis is the task of grouping a set of objects or very similar data in such a way that objects in same group or cluster are very similar to each other than to those in another groups or clusters. It is an unsupervised learning technique, which offers different views to inherent structure of a given dataset by dividing it into a many number of overlapping or disjoint groups. The different algorithm that we used in this paper to perform the cluster analysis of a particular given dataset is listed below.

4.1 Partition Based Clustering

It is based on the concept of iterative relocations of the data points from the given dataset between the clusters.

4.1.1 K-Means

The aim of this algorithm is to reduce objective function. Hers, the objective function that is considered is Square error function.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i - c_j \right\|^2$$

Where $\left\| x_i - c_j \right\|^2$ is the distance between the data point xi, and even the cluster points centroid cj.

Algorithm Steps:

- Consider a hepatitis dataset/data frame, load and pre-process the data
- Keep K points into the workspace as presented by the objects that has to be clustered. These are called the initial or starting group centroids.
- Here, the number of clusters is considered as 3.

- Closest centroid being identified and each object has been assigned to it.
- When all objects been assigned, the centroids is recalculated back again.
- Repetition is being done with Steps 2 and 3, till the centroids have no longer move.
- This gives out a separation of the objects into the groups from where the metric to be minimized should be calculated (Fig. 1).

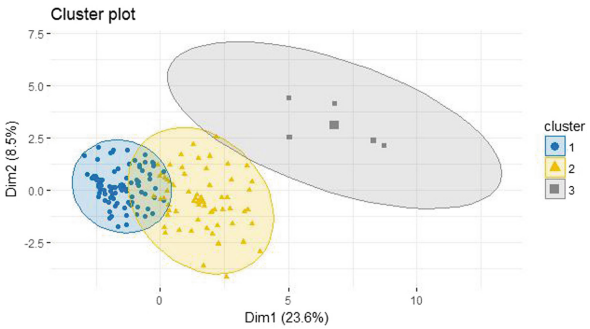

Fig. 1. K-means technique performed on Hepatitis data set in R studio.

4.1.2 K-Mediods (Partitioning Around Mediods)

K-Medoids algorithm is one of the partitioning clustering method that has been modified slightly from the K-Means algorithm. Both these algorithms, are particular meant to minimize the squared – error but the K-medoids is very much strong than the K-mean algorithm.

Here the data points are chosen as such to be medoids.

Algorithm steps:

- Load the dataset and pre-process the data
- Select k random points that considered as medoids from the given n data points of the Hepatitis dataset.
- Find the optimal number of clusters.
- Assign the data points as such to the closest medoid by using a distance matrix and visualize it using fviz function (Fig. 2).

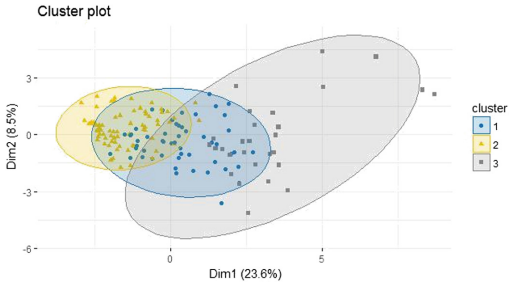

Fig. 2. K-medoids technique performed on hepatitis dataset using R studio

4.2 Hierarchy Based Clustering

This clustering basically deals with hierarchy of objects. Here we need not to pre-specify the number of clusters in this Clustering technique, like K-means approach. This clustering technique has been divided into two major types.

4.2.1 Agglomerative Clustering

This clustering technique is also known as AGNES, which is none other than Agglomerative Nesting. This clustering works as in bottom-up manner (Fig. 3).

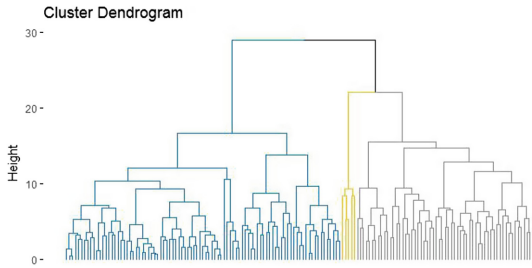

Fig. 3. Agglomerative clustering based on Hepatitis dataset in R studio

Algorithm Steps:

- Load and pre-process dataset then load the factoectra, nbclust, fpc packages..
- Assign each data object to a formed clusters such a way, that each object is assigned to one particular cluster.
- Find nearest pair of such clusters and combine them to form a new node, such that those are left out with N-1 clusters.
- Calculate distance between old and new clusters.
- Perform previous two steps till all the clusters have been clustered as one size.
- As we have N data objects, N clusters to be formed.
- At last, the data is visualized as a tree known as dendogram.

$$\text{d} = \sum_{i=1}^{n} |x_i - y_i| \quad \text{Manhattan Formula}$$

$$\text{d} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \text{Euclidean Formula}$$

4.2.2 Divisive Clustering

Divisive Clustering is just the opposite is Agglomerative algorithm. Divisive Clustering Approach is also known as Diana [3] (Fig. 4).

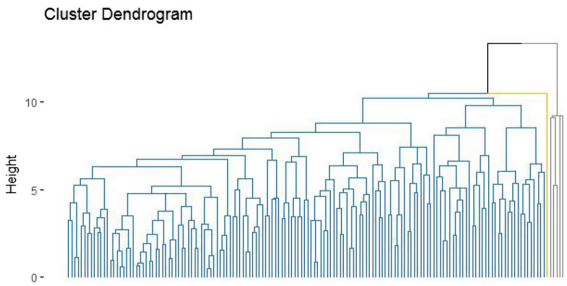

Fig. 4. Divisive clustering based on Hepatitis dataset in R studio

4.3 Fuzzy Clustering

Fuzzy is a method of clustering one particular piece of data is to belong to one or more clusters. It is based on the minimization of the objective function.

$$J_m = \sum_{I=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad 1 \leq m < \propto$$

Where m is a real number which is greater than 1, u is a degree of membership of xi, in the i^{th} dimensional data, cj is the centre dimension of the cluster.

Algorithm Steps:

- Load the dataset.
- Load the fanny function.
- At k – steps: Calculate the centres of the vectors c(k) = c(j) with U(k).

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_j}{\sum_{i=1}^{N} u_{ij}^{m}}$$

- Update the values of $U(K), U(K+1)$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m+1}}}$$

- If the $U(K+1) - U(K) < E$ then stop of the function, otherwise return back to Step 3.
- Visualize the data in the clustered format (Fig. 5).

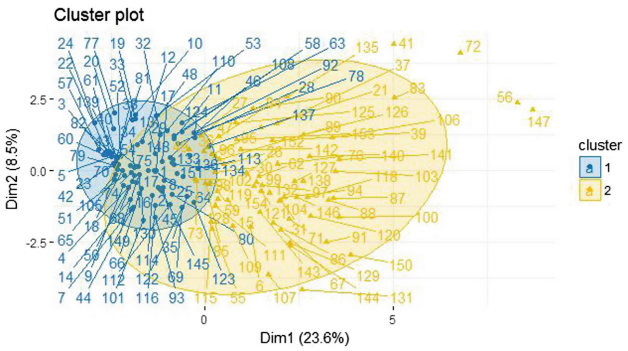

Fig. 5. Fuzzy clustering based on Hepatitis dataset in R studio

4.4 Model Based Clustering

The data will considered here is a mixture of two or more clusters.

Algorithm Steps:

- Load and pre-process the Hepatitis dataset.
- Install Mass, ggpubr, factoextra, mclust packages in library in R studio.
- Apply mclust function to cluster the data. Then visualize the data (Fig. 6).

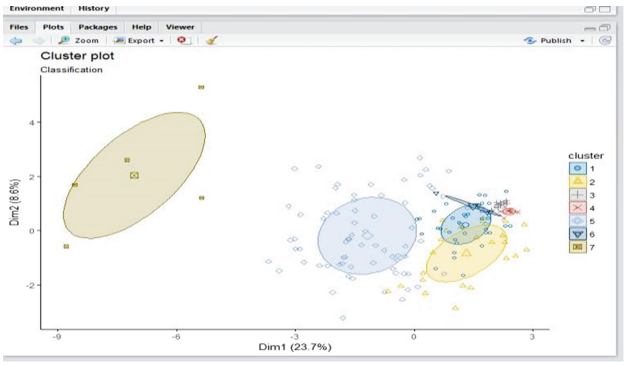

Fig. 6. Model based clustering based on Hepatitis dataset in R studio

5 Performance Analysis

5.1 Cluster Validation

Here the term of cluster validation is used here to evaluate and compare the goodness and accuracy of different clustering algorithms results. This Internal Cluster Validation, basically uses the internal information of all the clustering process to find out the effectiveness and goodness of a cluster structure without knowing the external

information. Internal measures results upon Compactness, separation and connectedness. Internal validation is done using Silhouette, Connectivity and Dunn Index.

$$\text{Index} = (x * \text{Separation})/(y * \text{Compactness})$$

Here x and y are the weights.

6 Results of Different Validation Techniques Using Dataset

See Figs. 7, 8 and 9.

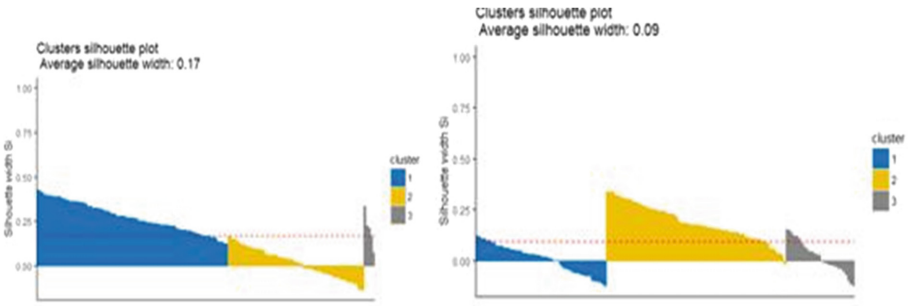

Fig. 7. K-means and K-medoids validations

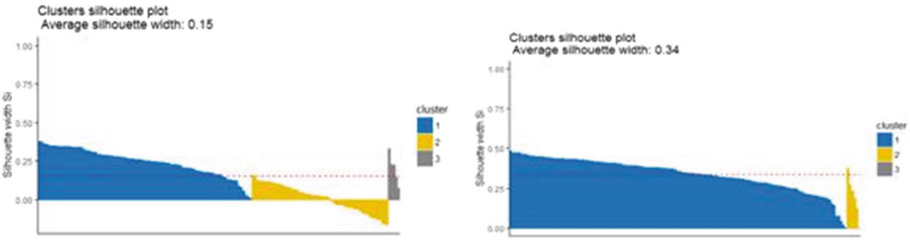

Fig. 8. Agglomerative and divisive validations

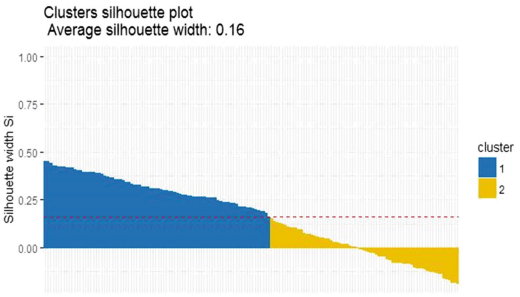

Fig. 9. Fuzzy validation

7 Choosing the Best Algorithm

Internal Validation of different clustering techniques results are listed here (Table 1).

Table 1. Comparision of clustering algorithms

	Connectivity	8.0996	59.2643
K-means	Dunn	0.6354	0.2907
	Silhouette	0.4395	0.1727
	Connectivity	8.0996	11.0286
K-medoids	Dunn	0.6354	0.6283
	Silhouette	0.4395	0.3358
	Connectivity	8.0996	11.0286
Diana	Dunn	0.6354	0.6283
	Silhouette	0.4395	0.3358
	Connectivity	75.3393	102.3099
Pam	Dunn	0.1061	0.2053
	Silhouette	0.1643	0.0942
	Connectivity	49.1099	NA
Fanny	Dunn	0.2004	NA
	Silhouette	0.1633	NA
	Connectivity	60.4056	NA
Model	Dunn	0.2138	0.14
	Silhouette	0.1298	−0.0289

8 Conclusion

This paper deals with defining few algorithms, and all those algorithms have been implemented and visualized in R studio. The clustering is done on hepatitis dataset. All the algorithms have been validated using internal measures and results have been displayed in the tabular format in terms of connectivity, Dunn, silhouette index. The measure has been considered for every algorithm and then compared overall to find out the best algorithm. As, per this we conclude that, the K-means is used for the large datasets and large number of clusters, Fuzzy clustering is not well suitable for the large number of clusters and also K-means have maximum dunn and silhouette index values when compare to all other algorithms.

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. **147**, 195–197 (1981)
2. May, P., Ehrlich, H.-C., Steinke, T.: ZIB structure prediction pipeline: composing a complex biological workflow through web services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006). https://doi.org/10.1007/11823285_121
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid ınformation services for distributed resource sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The physiology of the grid: an open grid services architecture for distributed systems ıntegration. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov
7. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 881–892 (2002)
8. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: 2010 IEEE 10th International Conference on Data Mining (ICDM), pp. 911–916. IEEE, December 2010
9. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., Wu, S.: Understanding and enhancement of internal clustering validation measures. IEEE Trans. Cybern. **43**(3), 982–994 (2013)