International Journal of Statistics and Applied Mathematics 2020; 5(4): 246-248

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2020; 5(4): 246-248
© 2020 Stats & Maths
www.mathsjournal.com
Received: 01-05-2020
Accepted: 05-06-2020

Avinash Navlani
School of Data Science and Forecasting, Devi Ahilya Vishwavidyalaya, New Delhi, Delhi, India

Dr. VB Gupta
School of Data Science and Forecasting, Devi Ahilya Vishwavidyalaya, New Delhi, Delhi, India

Corresponding Author:
Avinash Navlani
School of Data Science and Forecasting, Devi Ahilya Vishwavidyalaya, New Delhi, Delhi, India

Comparing clustering algorithms performance using multiple-objective functions

Avinash Navlani and Dr. VB Gupta

Abstract

Clustering is the bunching of the data into groups of identical objects. Here each bunch is known as a cluster, each object is identical to its objects of the same cluster and different from other clusters. In this paper, we are doing an experimental study for comparing clustering algorithms using multiple-objective functions. We have investigated K-means a Partitioning-based clustering, Hierarchical clustering, Spectral clustering, Gaussian Mixture Model Clustering, and Clustering using Hidden Markov Model. The performance of these methods was compared using multiple objective functions. Multiple objectives have two core objectives: Cluster Homogeneity and separation. These multiple objective functions will be a great help to discover robust clusters in a more efficient way.

Keywords: K-Means clustering, hierarchical clustering, spectral clustering, Gmm clustering, hmm clustering, cluster evaluation, multiple objective function

Introduction

Clustering is the task of grouping similar objects into the same category when separating objects, which are not similar. Clustering has been widely used since 40 years ago for solving various data mining problems. Data clustering that is known as an unsupervised classification technique is a very important field in machine learning that has applications in segmentation, summarization, and target detection (Jain and Dubes 1998) [6]. There are many different approaches introduced for clusterings such as Partitioning, Hierarchical, Density-Based, Probabilistic Model-Based, High-Dimensional, and Hidden Markov Model-Based Clustering. But still, there are some difficulties in cluster data based on evaluating different clusters and many studies are conducted and discussed the matter of cluster evaluation. In this paper, we have used homogeneity and separation measures used for Multi-objective function creation. These multiple objective functions can be applied to the clustered data to evaluate the performance of the clustering algorithm.

The paper is organized as follows, clustering algorithms discussed in section II, multiple objective functions in section III, Experimental results and discussion in section IV. Finally, the paper is concluded and future work in section V.

Clustering Algorithms

Based on the definition, clustering is a data mining technique that helps to organize objects to different categories that are called clusters. Clustering algorithms work in such a way that all objects belong to a cluster are very similar and they are from the same type and those in other clusters are from different types (Aggarwal and Reddy. 2013) [1]. Let's discuss the clustering algorithm that we experimented in this paper:

K-means Clustering

K-means is a partitioning based method. Partitioning-based methods divide data into multiple cluster partitions and then iteratively each data point is assigned to a cluster that has the closest distance to the centroid. In K-means, dataset, and the number of clusters K is given input to the algorithm. Here, data is divided into K partitions and data points assigned to the centroid of the closest partition.

K-means initiated with random cluster centers and iteratively changes the centers until there is no change in the cluster objects. K-means is sensitive to initial random cluster centroids and outliers. It generates spherically shaped clusters (Li and Wu 2012) [7].

Hierarchical Clustering

Hierarchical clustering usually is represented by a dendrogram that uses the similarity between two data points or two splits to create a tree-like cluster. Hierarchical clustering merges or splits the groups and builds a tree-like structure. This tree structure represents how groups are merged and split. Hierarchical clustering methods can be categorized into agglomerative and divisive hierarchical clustering based on merging or decomposition strategies such as bottom-up, top-down. In both types, a single cluster at the top and individual objects at the bottom (Han, *et al*. 2002) [5]. In our experiments, we are using agglomerative hierarchical clustering; it is a bottom-up approach that merges the microclusters into macro clusters at each intersection.

Spectral Clustering

The spectral clustering is a graph-based method used to group the unsupervised dataset. It captures the data into a weighted graph, reduces the dimension of the data using eigenvectors of similarity matrix and forms similar groups (Nadler and Galun 2006) [9]. It is easy to implement and can be capable of solving using linear algebra (Von Luxburg, Belkin and Bousquet 2008) [12]. It factorizes the Laplacian matrix which is generated from similarity matrix and partitions data into disjoint clusters with high within-cluster similarity and low between cluster similarity. Spectral clustering offers various applications in image processing, pattern recognition, speech recognition, data mining, and machine learning.

HMM Clustering

Hidden Markov Models (HMMs) are broadly accepted in the model sequential processes and capable of managing the multivariate data. HMM is also used in grouping objects. Still, we have very few papers in this area. In HMM clustering, we compute the pairwise distance and create a symmetric matrix of log-likelihood of each sequence by training an HMM. This LL matrix is used to cluster the data using hierarchical algorithms (Bicego, Murino and Figueiredo n.d.).

GMM Clustering

The Gaussian Mixture Model clustering is a probabilistic model. It is more flexible and generalized compared to k-means (McLachlan, Lee and Rathnayake 2019) [8]. GMM assumes that all the data points are a mixture of Gaussian distributions and maximizes the likelihood(Sahbi n.d.). GMM allows density estimation and uses information-theoretic penalties for model selection. Here, each data point belongs to each cluster with a specific probability (Guyeux, *et al*. 2019) [4].

Multiple Objective Functions

In this paper, we will assess the clustering performance using two multiple objective functions because a single objective is not enough to assess the performance of the clustering algorithm. For more robust clustering results we need to take both homogeneity and separation objectives. Here, the first function is for the minimization problem and the second function is for the maximization problem. Both functions are inspired by (Saha and Mitra 2014) [10] paper. Let's see an example of a minimization problem:

$$\text{Objective Function(Min.)} = \alpha * \text{DBI} + \left(\frac{1-\alpha}{2}\right) * \text{RI} + \left(\frac{1-\alpha}{2}\right) * \text{JI} \qquad (1)$$

Where α is the trade-off parameter between clustering quality and dissimilarity. Here, we have used the Davies-Bouldin Index (DBI) as internal evaluation parameters used to assess the cluster homogeneity and its lower value is preferable. Rand Index (RI), and Jaccard Index(JI) as an external evaluation parameter. Both explain the separation between clusters and lower values for both the measure is preferable. Let's see another objective function of the maximization problem:

$$\text{Objective Function(Max.)} = \alpha * \text{SC} + \left(\frac{1-\alpha}{2}\right) * \text{VM} + (1-\alpha) * \text{MI} \ (2)$$

Where α is the trade-off parameter between clustering quality and dissimilarity. In the above objective function, Here, we have used the Silhouette Coefficient (SC) as internal evaluation parameters used to explain the cluster homogeneity and higher value is preferable. V-Measure and Mutual Information (MI) score as an external evaluation parameter. Both used to explain the separation between clusters and the higher value of both measures preferable.

Experimental Results and Discussions

In this section, we evaluate and compare the performance of various clustering algorithms using multiple objective functions. Here the lower value of the minimum objective function and a large value for maximum objective function is considered as better clustering.

Datasets

In our experiments, we have used 7 datasets. Four of them are real-life datasets (Iris, Pima Indian diabetes, Liver patient, and Breast cancer) and 3 are synthetic datasets. The real-life datasets are taken from the UCI Repository (Dua and Graff 2017) [3].

Results of Clustering Algorithms

In our experiment, we applied five clustering algorithms such as K-means clustering, Hierarchical clustering, Spectral clustering, Gaussian Mixture Model Clustering, and HMM Clustering on four real-life datasets (Iris, Pima Indian diabetes, Liver patient, and Breast cancer) and three synthetic datasets. The results for minimization objective function are depicted for all the clustering algorithm for each data set and shown using heatmap in Figure-1.

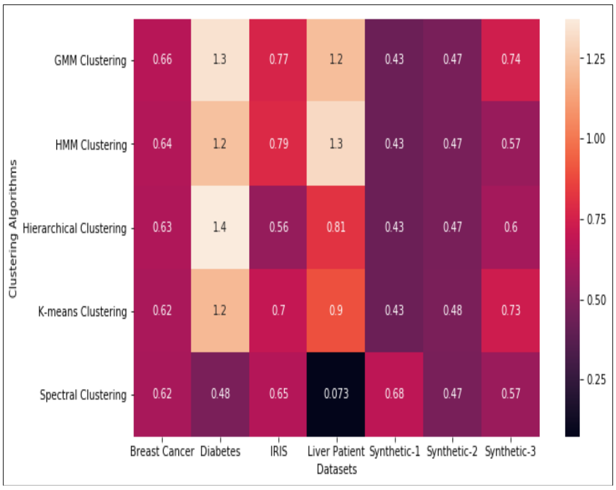

Fig 1: Minimization Objective Function Score

In the minimization objective function results, we can see that the spectral clustering is performing better on five datasets (Breast Cancer, Diabetes, Live patient, Synthetic-2 and Synthetic-3) and Hierarchical clustering is performing better on two datasets (Iris, and Synthetic-1). Also, you can see on Synthetic-1 and Synthetic-2 dataset most of the algorithms are giving the same score because these both have spherical well-separated clusters and the Synthetic-3 dataset has comparatively higher scores than other two synthetic datasets because this data has moon shape clusters. The results for maximization objective function are depicted for all the clustering algorithms for each data set and shown using a heatmap in Figure-2.

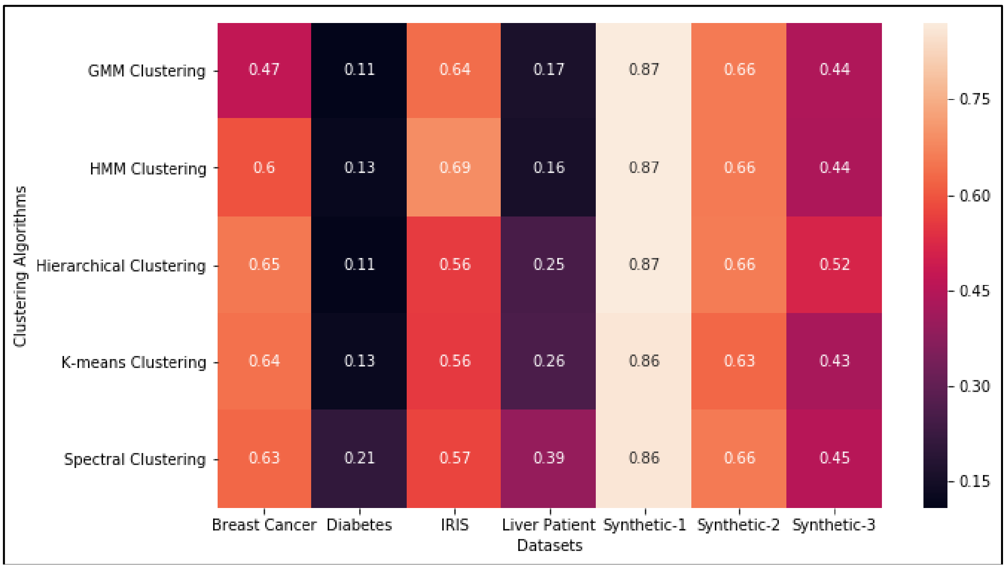

Fig 2: Maximization Objective Function Score

In the maximization objective function results, we can see that the spectral clustering is performing better on three datasets (Diabetes, Live patient, Synthetic-2) and Hierarchical clustering is performing better on four datasets (Breast cancer, Synthetic-1, Synthetic-2, and Synthetic-3). Also, you can see on Synthetic-1 and Synthetic-2 dataset most of the algorithms are giving the same score because these both have spherical well-separated clusters and the Synthetic-3 dataset has comparatively lower scores than other two synthetic datasets because this data has moon shape clusters.

Conclusion

In this paper, an experimental study of five well-known clustering algorithms such as K-means clustering, Hierarchical clustering, Spectral clustering, Gaussian Mixture Model Clustering, and Clustering using Hidden Markov Model was analyzed. The performance of clustering algorithms compared using maximize and minimize multiple objective functions. We have used four real-life datasets and three synthetic datasets for experimentation. In our results, spectral and hierarchical clustering performs better compared to other algorithms. Multi-objective function measures can be helpful to get a robust set of clusters. In future work, we will try to improve these objective functions and develop clustering algorithms based on these multiple objectives.

References

1. Aggarwal, Charu C, Chandan K Reddy. Data Clustering: Algorithms and Applications. CRC Press, 2013.
2. Bicego, Manuele, Vittorio Murino, Mário AT, Figueiredo ND. Similarity-Based Clustering of Sequences using Hidden Markov Models.
3. Dua, Dheeru, Casey Graff, 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/.
4. Guyeux, Christophe, Stéphane Chrétien, Gaby Bou Tayeh, Jacques Demerjian, Jacques Bahi. Introducing and Comparing Recent Clustering Methods for Massive Data Management in the Internet of Things. Journal of Sensor and Actuator Networks (MDPI AG). 2019; 8(4):56.
5. Han, Jiawei, Micheline Kamber, Fernando Berzal, Nicolás Marín. Data Mining: Concepts and Techniques. SIGMOD Record. 2002; 31(2):66-68.
6. Jain, Anil K, Richard C Dubes. Algorithms for Clustering Data. Prentice-Hall, Inc, 1998.
7. Li, Youguo, Haiyan Wu. A Clustering Method Based on K-Means Algorithm. Physics Procedia (Elsevier BV). 2012; 25:1104-1109.
8. McLachlan, Geoffrey J, Sharon X Lee, Suren I Rathnayake. Finite Mixture Models. Annual Review of Statistics and Its Application (Annual Reviews). 2019; 6(1):355-378.
9. Nadler, Boaz, Meirav Galun. Fundamental Limitations of Spectral Clustering. Proceedings of the 19th International Conference on Neural Information Processing Systems, 2006, 1017-1024.
10. Saha, Moumita, Pabitra Mitra. VLGAAC: Variable length genetic algorithm based alternative clustering. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, 2014, 194-202.
11. Sahbi, Hichem ND. A Particular Gaussian Mixture Model for Clustering.
12. Von Luxburg, Ulrike, Mikhail Belkin, Olivier Bousquet. Consistency of Spectral Clustering. The Annals of Statistics. 2008; 36(2):555-586.