Hierarchical and EM algorithm based clustering methods: A comparative review

Bibin Babu
MAI, Faculty of Computer Science and
Business Information System

University of Applied Sciences
Würzburg-Schweinfurt
Germany
bibin.babu@student.fhws.de

***Abstract* — Numerous real-world systems may be analyzed on the basis of pattern recognition problems, making correct application (and comprehension) of machine learning approaches crucial. There is no agreement on which classification algorithms are best for a specific dataset, despite the fact that several have been offered. As a result, it's critical to thoroughly compare approaches in a variety of settings. In this context, a comprehensive comparison of two well-known clustering methods conducted: Clustering using expectation maximization algorithm and hierarchical clustering which are accessible in the R language, both of which assume normally distributed data. To accommodate for the wide range of potential data changes, simulated datasets with a variety of configurable attributes (number of classes, separation between classes, etc.) were used to assess the robustness of clustering algorithms to parameter setting.**

Keywords — Hierarchical clustering, Expectation–maximization algorithm, Jaccard Index, Adjusted Rand Index, Fowlkes Mallows Index, Normalized Mutual Information.

I INTRODUCTION

Clustering algorithms are used to find groupings of items or clusters that are more similar to one another than to other clusters. Such a method of data analysis is strongly linked to the work of developing a data model, which entails establishing a reduced collection of attributes that may be used to offer an intelligible explanation of key features of a dataset. Clustering algorithms are generally more demanding than supervised learning, and they reveal more information about complicated data.

The advantage of simulated data is the ability to collect an infinite number of samples and to modify any of the aforementioned aspects of a dataset in a systematic manner. Such characteristics enable the clustering algorithms to be thoroughly and rigorously assessed in a wide range of scenarios, as well as evaluating the performance's sensitivity to slight changes in the data. The similarity between the known labels of the items and those discovered by the algorithm is used to measure performance. In which the Jaccard index [2], the Adjusted Rand index [3], the Fowlkes-Mallows index [4], and the Normalized mutual information [5] are all compared. To measure the performance of the clustering methods, a modified version of the approach provided by [6] was employed to construct 400 different datasets and go over the technique that was utilized and the parameters that were used to generate the data. datasets created using a comprehensive process reported earlier in [12]. The amount of features, classes, objects per class, and average distance between classes may all be modified consistently across datasets. Here the program is implemened in R language. The prominence of the R language in the data mining area, as well as the well-established clustering packages it provides, drove this decision. The definition of a good technique for setting parameter values has long been a difficulty in machine learning [7]. In theory, an optimization process (e.g., simulated annealing [8] or genetic algorithms algorithms [9]) may be used to identify the parameter configuration that gives a particular program the optimum performance. However, there are two key drawbacks to this strategy. First, overfitting can occur when parameters are adjusted to a specific dataset [10]. That is, when fresh data is reviewed, the precise parameters established to produce optimal performance may result in worse performance. Second, due to the time complexity of many algorithms paired with their normally high number of parameters, parameter optimization may be impractical in some circumstances. Many scholars eventually fall to employing classifier or clustering methods with the software's default parameters. As a result, time and effort will be necessary to evaluate and compare the performance of clustering algorithms in both optimization and default scenarios. We will look at some exemplary instances of algorithms used in the literature [7, 11] in the following sections.

II RELATED WORKS

in [13] used three datasets of documents to conduct a comparative comparison of clustering algorithms in the context of a text-independent speaker verification task. Clustering methods focusing on minimizing a distance-based objective function and a Gaussian models-based approach were also investigated. K-means, random swap, expectation-maximization, hierarchical clustering, self-organized maps (SOM), and fuzzy c-means were all compared. The authors discovered that the model order, which indicates the number of centroid or Gaussian components (for Gaussian models-based techniques) examined, is the most essential element for the algorithms' effectiveness. Overall, clustering algorithms that focused on decreasing a distance-based objective function had similar recognition accuracy. Hierarchical techniques were less accurate than the EM algorithm method when the number of clusters was minimal.

III HIERARCHICAL AND EM ALGORITHM CLUSTERING METHODS

In hierarchical approaches, which take into account the linkage between data points can be split into two: agglomerative and divisive [14]. Initially, each object in an agglomerative hierarchical clustering method belongs to its own distinct cluster. After then, groups are merged in consecutive iterations till stop conditions are met. A divisive hierarchical clustering approach, on the other hand, starts with all objects in a single cluster and separates them into clusters after several repetitions. The stats and cluster packages are the two primary tools in the R language that contain methods for conducting hierarchical clustering. We'll look at the agnes routine from the cluster package, which executes the algorithm presented by Kaufman and Rousseeuw [15]. In agnes, there are four main linking criteria: single linkage, complete linkage, Ward's approach, and weighted average linkage [16].

Model-based approaches may be thought of as a broad framework for calculating the maximum likelihood of an underlying distribution's parameters for a given dataset. The expectation-maximization (EM) technique is a well-known example of model-based methodologies. The most prevalent concept is that the data from each class can be described using multivariate normal distributions, and that the distribution seen for the entire data can thus be viewed as a combination of these normal distributions. The most probable parameters of the normal distributions of each class are then determined using a maximum likelihood technique. When the dataset is incomplete, the EM technique for clustering is very useful [17, 18]. But it may greatly dependent on the initial settings [19]. Furthermore, the method may not be able to detect extremely tiny clusters [20, 21]. In the R program mclust package [22, 23] provides iterative EM (Expectation-Maximization) techniques for maximum likelihood estimation using parameterized Gaussian mixture models utilizing iterative EM methods. The various stages of an EM iteration are implemented by the functions estep and mstep.

IV MATERIALS AND METHODS

IV.A Datasets

To properly compare clustering algorithms, you'll need a reliable simulated data creation approach that can generate a range of datasets. We use an approach based on Hirschberger et alearlier's work [6] for such a task. The process may be used to create samples that are typically distributed and divided into C classes based on F characteristics. A covariance matrix Ri of dimension F × F is constructed for each class i in the dataset, and this matrix is utilized to generate Ne objects for the classes. This means that for each created class, pairs of characteristics might have different correlations. The created class values are then divided by α and translated by si, where si is a random variable characterized by a uniform random distribution specified in the interval [1, 1], and si is a random variable described by a uniform random distribution defined in the interval [1, 1]. The predicted distances between classes are related with parameter α . Depending on the amount of objects and characteristics in the dataset, such distances can have varying effects on clustering. DB2F, DB10F, DB50F, and DB200F are the acronyms for datasets with 2, 10, 50, and 200 features, respectively. Such datasets are made up of 50 items for each class (i.e., Ne = 50) and C = 2, 10, and 50 elements for each class (i.e., C = 2, 10, 50). In certain circumstances, the number of classes examined for the dataset is also shown. The dataset DB2C10F, for example, has two classes, ten features, and 50 entries per class.

IV.B Experiments and Evaluation on the performance of clustering algorithms

The Jaccard Index (J) [2], Adjusted Rand Index (ARI) [3], Fowlkes Mallows Index (FM) [4], and Normalized Mutual Information (NMI) [5] are the most conventional external indices used here. The following concepts are taken into account while defining the cluster external index. Let U = u1, u2,...uR reflect the dataset's original partition, with ui denoting a subset of the items associated with cluster i. Let V = v1, v2,...vC represent the partition discovered by a cluster method. In both U and V, we express the number of pairs of items that are put in the same group as a may be calculated mathematically by

$$a = \sum_{i,j} \binom{n_{ij}}{2}, \tag{1}$$

where nij is the number of items in both the ui and vj subsets.

Let b denote the number of pairings of items in U that belong to the same group but not in V, i.e.

$$b = \sum_{i} \binom{n_{i.}}{2} - \sum_{i,j} \binom{n_{ij}}{2}, \tag{2}$$

where ni is equal to sum of nij in j. Let c denote the number of pairs of objects in U that belong to distinct groups but the same group in V, which can be expressed as

$$c = \sum_{j} \binom{n_{.j}}{2} - \sum_{i,j} \binom{n_{ij}}{2}, \tag{3}$$

where nj is equal to I nij. Based on a, b, and c, the Jaccard Index (J), Adjusted Rand Index (ARI), and Fowlkes Mallows (FM) indexes may be defined:

$$J = \frac{a}{a+b+c}, \tag{4}$$

$$\mathrm{ARI} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{1/2\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}, \tag{5}$$

$$\mathrm{FM} = a\frac{\sqrt{a+b}\sqrt{a+c}}{(a+b)(a+c)}. \tag{6}$$

We also utilize the normalized mutual information (NMI) metric as a quality metric since it quantifies the mutual dependency between two random variables using well-known information theory ideas. In [27] we can find the definition of the NMI metric detaily.

$$\mathrm{NMI}(C,T) = \frac{I(C,T)}{\sqrt{[H(C),H(T)]}}. \tag{7}$$

where C is the random variable indicating the point cluster assignments, and T is the random variable indicating the point's underlying class labels. The mutual information between the random variables C and T is I(C, T) = H(C) H(C| T). The Shannon entropy of C is H(C). The conditional entropy of C given T is H(C|T).

V RESULTS AND DISCUSSION

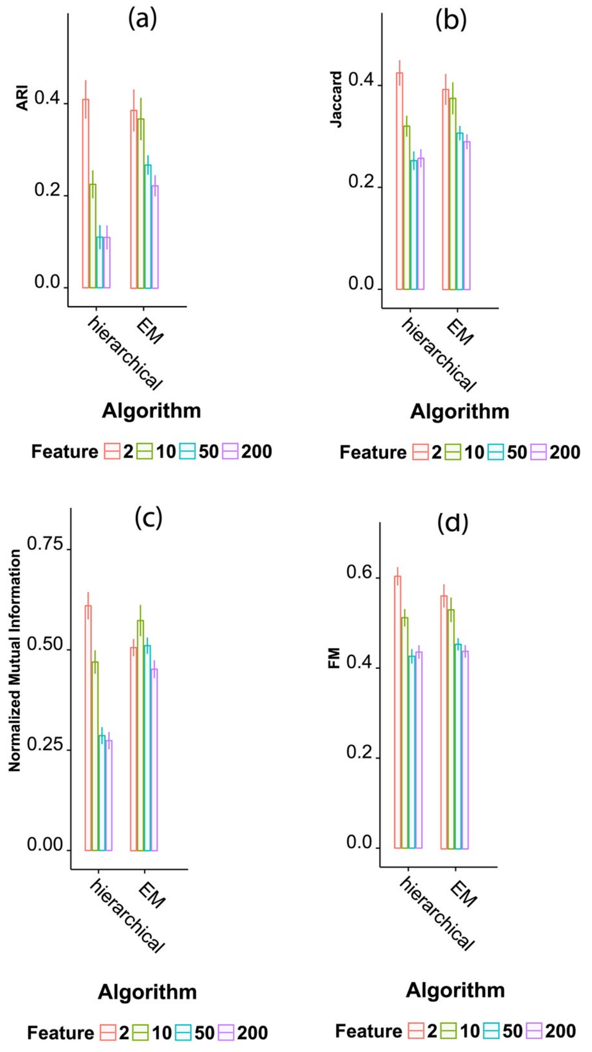

Fig 1. The average performance of the two clustering techniques based on the amount of characteristics in the dataset.

The acquired results for the four examined performance measures are shown in Figure 1. The amount of features in the dataset appears to have a significant impact on the hierarchical technique. In particular, the hierarchical technique was less accurate when 50 and 200 characteristics were used. It is indeed evident that the algorithms' output for two features appears to be close.

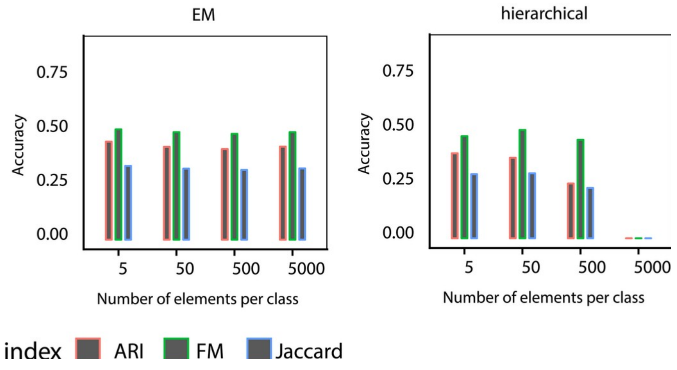

Fig 2. Performance of the algorithms when the number of elements by class correspond to Ne = 5, 50, 500, 5000.

The size of the data has an influence on the clustering quality in most clustering techniques. We analyzed a situation with a lot of circumstances in order to measure this effect. Datasets were built using F = 5, C = 10, and Ne = 5, 50, 500, 5000 instances per class. DB10C5F will be used to refer to this dataset. The size of the dataset appears to have almost no effect on the accuracy of the EM algorithms, as shown in Figure 2. Because of the quantity of memory required by these techniques, the accuracy of hierarchy could not be measured when 5000 instances per class were employed. We can also observe that when the number of entries per class grows, the accuracy of the ARI and Jaccard indexes decreases dramatically.

VI CONCLUSIONS

The choice of several distinct methodologies, parameters, and performance criteria is required when clustering data, which has implications for many real-world issues [24, 25, 26]. As a result, it is a challenging undertaking that has drawn a lot of attention to analyze the benefits and drawbacks of clustering techniques. Here, we focused on a comparing clustering methods using hierarchical and EM algorithm methods thorough a process of producing a wide variety of heterogeneous datasets with well specified attributes such the distances between classes and correlations between features, which used on 400 simulated datasets using R language program. Both the EM and the hierarchical algorithms demonstrated a noticeable gain in performance for datasets with two classes and 10 features. When number of features and number of elements per class increases clustering using EM algorithm stays almost stable. This shows that several parameter choices for this method can provide outcomes that are superior than those produced by the default setting. While clustering using hierarchical method shows reduction in the Index values. Future expansions of this study might compare these algorithms by analysing one-dimensionally and multi-dimensionally.

REFERENCES

1 Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. (2019) Clustering algorithms: A comparative approach. PLoS ONE 14(1): e0210236. https://doi.org/10.1371/journal.pone.0210236

2 Jaccard P. Nouvelles recherches sur la distribution florale. Bulletin de la Sociète Vaudense des Sciences Naturelles. 1908;44:223–270.

3 Lawrence H, Arabie P. Comparing partitions. Journal of Classification. 1985;2(1):193–218.

4 Fowlkes E B, Mallows C L. A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association. 1983;78(383):553–569.

5 Strehl A, Ghosh J, Cardie C. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research. 2002;3:583–617.

6 Hirschberger M, Qi Y, Steuer RE. Randomly generating portfolio-selection covariance matrices with specified distributional characteristics. European Journal of Operational Research. 2007;177(3):1610–1625.

7 Berkhin P. In: Kogan J, Nicholas C, Teboulle M, editors. A Survey of Clustering Data Mining Techniques. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 25–71.

8 Hwang CR. Simulated annealing: theory and applications. Acta Applicandae Mathematicae. 1988;12(1):108–111.

9 Goldberg DE, Holland JH. Genetic algorithms and machine learning. Machine learning. 1988;3(2):95–99.

10 Hawkins DM. The problem of overfitting. Journal of chemical information and computer sciences. 2004;44(1):1–12. pmid:14741005

11 Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM computing surveys. 1999;31(3):264–323.

12 Amancio DR, Comin CH, Casanova D, Travieso G, Bruno OM, Rodrigues FA, et al. A systematic comparison of supervised classifiers. PloS one. 2014;9(4):e94137. Pmid:24763312

13 Kinnunen T, Sidoroff I, Tuononen M, Fränti P. Comparison of clustering methods: A case study of text-independent speaker modeling. Pattern Recognition Letters. 2011;32(13):1604–1617.

14 Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010; 31(8):651–666. https://doi.org/10.1016/j.patrec.2009.09.011

15 Kaufman L, Rousseeuw PJ. Finding Groups in Data: an introduction to cluster analysis. John Wiley & Sons; 1990.

16 Lance GN, Williams WT. A general theory of classificatory sorting strategies II. Clustering systems. The computer journal. 1967; 10(3):271–277. https://doi.org/10.1093/comjnl/10.3.271

17 Redner R, Walker H. Mixture densities, maximum likelihood and the em algorithm. SIAM Review.1984; 26(6).

18 Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B. 1977; 39(6).

19 Aggarwal CC, Reddy CK. Data Clustering: Algorithms and Applications. vol. 2. 1st ed. Chapman & Hall/CRC; 2013.

20 Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal. 1998; 41(8):578–588. https://doi.org/10.1093/comjnl/41.8.578.

21 Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association. 2002; 97(458):611–631. https://doi.org/10.1198/016214502760047131.

22 Fraley C, Raftery AE. MCLUST: Software for model-based cluster analysis. Journal of Classification.1999; 16(2):297–306. https://doi.org/10.1007/s003579900058.

23 Fraley C, Raftery EA. Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST”. Journal of Classification. 2003; 20(2):263–286. https://doi.org/10.1007/s00357-003-0015-3.

24 Raykov YP, Boukouvalas A, Baig F, Little MA. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. PLoS ONE. 2016; 11(9):1–28. https://doi.org/10.1371/journal.pone.0162259.

25 Arruda GF, Costa LF, Rodrigues FA. A complex networks approach for data clustering. Physica A: Statistical Mechanics and its Applications. 2012; 391(23):6174—6183. https://doi.org/10.1016/j.physa.2012.07.007.

26 Benaim M. A Stochastic Model of Neural Network for Unsupervised Learning. Europhysics Letters.1992; 19(3):241. https://doi.org/10.1209/0295-5075/19/3/015.

27 Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research. 2002; 3:583–617.