Comparisons between Automatic and Non-Automatic Clustering Algorithms

Jassim T. Sarsoh
Kadhem M. Hashim
Firas S. Miften
Thi-Qar University/ Education College of Pure Science/ Computer Science Dept

Abstract

This paper presents a comparative study between two famous types of clustering algorithms. These types are the automatic and non-automatic clustering algorithms.

The comparisons concerned some different criteria such as: dataset size, clusters number, execution time, results quality and accuracy. An effective automatic clustering algorithm is chosen as a sample for the automatic clustering techniques, while the well-known partitional K-Means clustering algorithm is taken as a sample for the non-automatic clustering techniques. The two chosen algorithms are implemented on the same database (ORL) concerning the human face images. Some conclusions are extracted to the performance of this implementation. MATLAB version (R2010a) is used to achieve the purpose of this paper.

Keywords: Data clustering, automatic clustering, non-automatic clustering, neighborhood, K-Means clustering.

مقارنة بين خوارزميات العنقده الآلية وخوارزميات العنقده غير الآلية

جاسم طعمه سرسوح
كاظم مهدي هاشم
فراس صبار مفتن
جامعة ذي قار/ كلية التربية للعلوم الصرفة / قسم علوم الحاسبات

المستخلص:

البحث الحالي يعرض دراسة مقارنة بين نوعين من خوارزميات العنقده، هذان النوعان هما: خوارزميات العنقده الآلية وخوارزميات العنقده غير الآلية. المقارنات المستخدمة في هذا البحث تخص بعض المعايير والضوابط مثل حجم البيانات، عدد العناقيد، زمن المعالجة، دقة النتائج. اختيرت خوارزمية عنقده ذاتية كفوءه كنموذج لخوارزميات العنقده الآلية بينما اختيرت خوارزمية K-means كنموذج لخوارزميات العنقده غير الآلية. نفذت الخوارزميتان المختارتان على نفس قاعدة البيانات وهي ORL التي تخص الوجوه البشرية وتم استخراج بعض المقارنات والاستنتاجات المفيدة بعد تنفيذ الخوارزميتين على قاعدة البيانات المذكورة أنفا.استخدمت لغة ماتلاب الإصدار (R2010a) لانجاز الهدف من البحث.

كلمات مفتاحيه : عنقده البيانات، العنقده الآلية، العنقده غير الآلية، الجيران، طريقة العنقده K-means

1. Introduction

Clustering is a division of data into different groups of similar objects. Each group, called cluster, consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups. Representing data by few clusters necessarily losses certain fine details, but achieves simplification and good interpretation for these data [1].

In general, comparisons among many algorithms in any scientific field must take into account some conditions and tools to get good comparisons results. In fact, the same conditions must be used for the chosen algorithms such as: criteria of comparisons, programming language, dataset, and programmer.

The objective of the comparisons is to highlight the strength and weakness of each algorithms compared to the others w.r.t each chosen factor of comparisons. This paper is investigated to compare between two types of clustering algorithms, those types are automatic and non-automatic clustering algorithms. The comparisons concern some different parameters such as: data size, clusters number, execution time, and percentage of the results success. We chosen an effective automatic clustering algorithms proposed in [2] as a sample of automatic clustering methods, and the famous partitional K-means clustering algorithm in [3] as a sample of non-automatic clustering methods.

The two chosen algorithms are implemented on the same dataset (ORL database) that includes the human face images for (60) persons and each person has (10) different human face images [4].

2-Clustering algorithms

Clustering algorithms partition data into a certain number of clusters (groups, subsets, or categories). There is no universally agreed upon definition. Most researchers describe a cluster by considering the internal homogeneity and the external separation such that patterns in the same cluster should be similar to each others, while patterns in different clusters should not. Different starting points and criteria lead to different taxonomies of clustering algorithms such as: hierarchical clustering, partitional clustering, Squared Error-based clustering, Mixture Densities -Based clustering, Graph Theory-Based clustering, Neural Network-Based clustering and Kernel_Based clustering. Some of these algorithms may be automatic, while the others may be non-automatic. There is no clustering algorithm that can be universally used to solve all types of problems [5,6,7].

Usually, clustering algorithms are designed with some assumptions and some types of biases. In this sense, it is not accurate to say best in the context of clustering algorithms, although some comparisons are possible. These comparisons are mostly based on some specific applications, under certain conditions, and the results may become quite different if the conditions change [5,6,7].

This paper aims to compare between automatic and non-automatic clustering algorithms by useing ORL database of humane face images. The comparisons based on some factors such as: data size, clusters number, execution time, and the percentage of results accuracy.

The following subsections (2.1 and 2.2) describe the principle ideas of automatic and none-automatic clustering techniques.

2.1. Automatic clustering techniques:

In this type of clustering, the number of clusters is not given a priori, and it is automatically determinated by the used clustering algorithms. The resulted clusters number is often closed to exact the number of grouping concerned the real structure of the studied dataset.

Jassim T. Sarsoh proposed an effective automatic clustering algorithm to group (cluster) the human face images by using the effect of facial segments features [2]. This algorithm is chosen as a sample of the automatic clustering techniques. The main idea of this algorithm is as follows:

•Determine the **adaptive neighbors** for each individual of the studied dataset, the **adaptive neighbors** depend on the chosen **threshold.**

•Compute the density of each human face image as follows

Density(x)=Cardinal(Adaptive _ neighbors(x))

Where **Density** is a vector of the number of adaptive neighbors for each element of the studied dataset, and **Density**(x) is the number of the adaptive neighbors of the individual **x**. let **V** is a vector containing the studied dataset.

- Sort the elements of the vector **Density** in descending order, and swap the corresponding face images in vector V according to the result of this sorting. The **adaptive neighbors**, will be also be swapped.
- The first element in vector **V** must construct (create) the first cluster since it has the largest number of **adaptive neighbors** in **Density**. All the **adaptive neighbors** of the first element in **V** must be located in this cluster.
- Therefore, the second element in **V** whose position corresponds the second element in **Density** must be taken as clustering candidate.
- If (this candidate has been assigned to any existed cluster) then
  all its adaptive neighbors must be located in that cluster.
  Else
  This candidate will construct another new cluster and all its **adaptive neighbors** must locate in this new cluster.
- The process will continue until the last element in **V** has been clustered in its corresponding cluster.

2.2. Non-Automatic clustering techniques:

In this type of clustering, the number of clusters must be given a priori by the programmer. The K-means clustering algorithm is a sample of this type. The accuracy of the obtained results depends on the predicted number of clusters chosen by the user when this algorithm is implemented on real dataset.

The pseudo code of K-Means clustering algorithm was found in [3]. This code will be as follows:

1. Choose **K** cluster centers to coincide with K randomly chosen parameters.
2. Assign each pattern of the studied dataset to the closet cluster center.
3. Recompute the cluster centers using the current cluster memberships.
4. If a convergence criterion is not met, go to step (2).

Typical convergence criteria are: no reassignment of patterns to new cluster centers, or minimal decrease in square error .

3. Related works

In the literature, some papers were found concerning the comparison among many clustering algorithms. The following are samples of those papers.

- Comparisons among four clustering algorithms were presented in [1]. Those algorithms are: k-means clustering, hierarchical clustering, self _ organizing map (SOM), and Expectation Maximization (EM) clustering. The chosen algorithms were implemented on some simple random and non-random datasets chosen from the web sites. As consequence, the partitional algorithm (K-means and EM) are recommended for huge datasets, while hierarchical clustering algorithms are recommended for small datasets. Hierarchical and SOM algorithms give better results compared to K-means and EM algorithms when choosing random datasets and vice versa.

- A comparison study between various fuzzy clustering algorithms was appeared in [8]. It concerned comparison between two famous fuzzy clustering techniques: fuzzy C-means (FCM) clustering algorithm and subractive clustering algorithm. High non-linear functions were modeled and the comparisons were made according to the capabilities of modeling. General conclusions indicate that number of clusters yields an improvement in the validity index value. The optimal modeling results were obtained when the validity indices are on their optimal values. The models generated from the subtractive algorithm are always more accurate than those generated using (FCM) clustering algorithms.

- Comparisons among the clustering algorithms (single linkage, complete linkage, average linkage, and ward hierarchical agglomerative for documents clustering and retrieval were shown in [9]. It was found that the average linkage clustering algorithm is the most suitable for documents clusters purposes.

In spite of some common features between our approach of comparisons w.r.t. other. Our approach is different from them by using an automatic and non-automatic clustering techniques. Besides we use complex dataset (human face images in the ORL database). In fact, the other researchers used only simple random and non random points in the plan or some selected vectors in the space, and some of these data were found in certain web sites.

4. Comparison criteria

It important to determine the criteria (factors) with which the comparisons among the algorithms must be achieved. In fact, the comparisons will determine the effect of each criterion on each chosen algorithm. We can conclude that an algorithm is the best among some studied algorithms w.r.t to certain criterion if the performance of that algorithm is the best. In this paper we proposed the following criteria: datasize, clusters numbers, execution time, and the percentage of the success results. We will notice the effect of each criterion in the comparisons approach among the following clustering algorithms. The automatic clustering algorithm chosen from [2] and the non-automatic clustering algorithm (K-means) chosen from [5]. As a result, some conclusions will be extracted from this comparative study.

5. Implementation

5.1. Experiment Results

Each of the two algorithms is implemented on the same dataset. This dataset is the ORL database which includes (60) persons, and for each person (10) different face images [4]. The implementation has been achieved by using matlab (ver. R2010a). This implementation has been processed as follows:

1- Each of the two algorithms is firstly implemented on (100) face images that concerned (10) persons of the ORL database. Then the two algorithms are implemented on the total data of the ORL database (600 face images for 60 persons).

2- For the automatic clustering algorithm, the user firstly choose a constant threshold which leads to calculate the adaptive threshold. This adaptive threshold is used to determine the adaptive neighbors which leads to give an optimal clustering results.

3- For K-means clustering algorithm, the user choose the number of clusters (K) a priori.

4- Figure [1] shows a sample of the automatic clustering algorithm results, while figure [2] shows a sample of the non-automatic K-means clustering algorithm results.

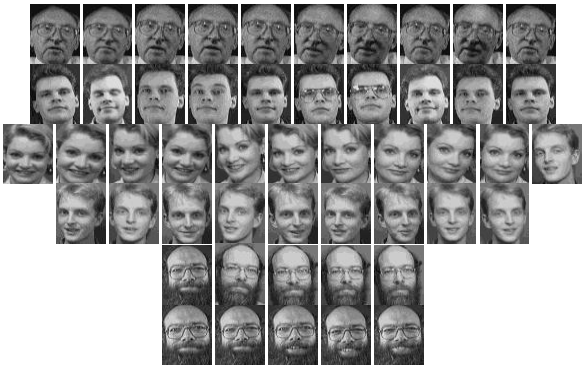

Figure (1): sample of the results for the automatic – clustering algorithm

Figure (2): sample of the results for the K-means algorithm

5- Table (1) shows the comparative results for the two algorithms, where each algorithm is firstly implemented on (100) face images concerned (10) persons, and secondly each of the two algorithms is implemented on (600) face images concerned (60) persons.

Table (1) : The comparisons between the result of the two algorithms.

The algorithm	Dataset size	Real Clusters	Obtained Clusters number	Execute time (seconds)	Success percentage
Automatic clustering algorithm	100	10	8	0.0972	71%
		10	10	0.0996	87%
		10	13	0.1053	98%
	600	60	47	0.6210	67%
		60	60	0.6320	82%
		60	75	0.7188	91%
Non-Automatic K-means clustering algorithm	100	10	8	0.8815	69%
		10	10	0.8494	82%
		10	13	0.9741	87%
	600	60	47	24.7159	61%
		60	60	21.3850	73%
		60	75	26.7553	79%

5.2. Discussions and Conclusions

1. For the automatic clustering, the number of the clusters is determinated automatically by this algorithm, while for the K-means clustering algorithm, the number of the clusters is given a priori by the user. As consequence, the automatic clustering algorithm is better than the K-means clustering algorithm because the first algorithm gives clustering results which simulate the real structure of the studied dataset.

2. We notice that the quality of the obtained results depends on the chosen values for the threshold in the automatic clustering algorithm, and on the chosen value for (K) in the K-means algorithm. In fact, choosing large value for threshold in automatic algorithm will cause to put the face images of two persons or more in the same cluster, while choosing small value for this threshold, will cause to put the face images of one person in two clusters, or more. Conversely, for the K-means algorithm, choosing small value to K will merge the face images of more than one persons in the same cluster, while choosing large value to K, will divide the face image for one person into many clusters.

3. Regarding the execution time criterion, Table(1) shows that the automatic clustering algorithm is usually better than K-mean algorithm.

4. Concerning the success percentage criterion, Table(1) shows that the automatic clustering algorithm gives always better results.

5. Finally, for the dataset size criterion, the automatic clustering algorithm is better than the K-means algorithm for the processing of the huge datasets.

6. References

[1] Osama Abu Abbas, "**Comparisons Between Data Clustering Algorithms**", The International Arab Journal of Information Technology, Vol.5, No.3, 2008.

[2] Jassim T. Sarsoh, "**Effect of Facial Segments Features on Human Face Classification**", Jornal of Basrah-Researches / Sciences Vol.34, N0.1, 2007.

[3] Jain A. K., Murty M. N., and Flyun P. J., "**Data Clustering: A Review**", ACM, Computing Surveys, Vol.31, No.31, 1995.

[4] http://www.machinelearning.ru/wiki/index.php?titel=The ORL_Database_of_Faces.

[5] Rui Xu and Donald C. Wunch,"**Survey of Neural Network**", Vol.16, No.3, 2005.

[6] Jain A., and Dubes R., "**Algorithms For Clustering Data**", Neural Computer Survey., Vol.37, No.12, 1989.

[7] Everitt B., Landau S., and Leece M., "**Cluster Analysis**", London, Arnold, 2001.

[8] Bataineh K. M., Naji M., and Saqer M., "**A Comparison Study Between Various Fuzzy Clustering Algorithms**", Journal of Mechanical and Industrial Engineering, Vol.5, No.4, P 335-343, 2011.

[9] El-Hamdouchi, and Willett P.," **Comparison of Hierarchic Agglomerative Clustering Methods**", The computer Journal Vol.32, No.3, 1989.