International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 4, Issue 1, 2013, pp119-123
http://bipublication.com

PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS FOR CHARACTER RECOGNITION USING WEKA TOOL

Sunila Godara and Ritu Yadav

Department of Computer Science and Engineering, GJUS&T, Haryana, India
*Corresponding author: Email: sunilagodara@gmail.com, ryadav1986@gmail.com

[Received-30/09/2012, Accepted-16/01/2013]

ABSTRACT:

Clustering is an unsupervised classification that is the partitioning of a data set in a set of meaningful subsets. Each object in dataset shares some common property- often proximity according to some defined distance measure. Among various types of clustering techniques, K-Means, Hierarchical and Make Density Based clustering are the most popular algorithms. Clustering Techniques are very useful in Character Recognition for automatically recognize the characters. In this paper we applied K-Means, Density Based and Hierarchical algorithms for clustering of Letter Image Recognition and Multi-Feature Digit data sets using WEKA machine learning tool. WEKA is a popular tool for machine learning which was written in java. The WEKA provides a collection of visualization tools and algorithms for data analysis and predictive modeling through a graphical user interface. Experimental results on Character Recognition data show that the k-means algorithms can make cluster in minimum time, and have good performance and the clustered accuracy is more than others two algorithms.

Keywords: Clustering, K-means, Hierarchical, Make Density Based Clustering, Weka

[I]. INTRODUCTION

Clustering is a type of categorization imposed rules on a group of data points or objects. A broad definition of clustering could be "the process of categorizing a finite number of data points into groups where all members in the group are similar in some manner". As a result, a cluster is an aggregation of objects. All data points in the same cluster have common properties (e.g. distance) which are different to the data points lying in other clusters.

A Character recognition System is a step towards the automation process, which requires in many fields, including image processing, office automation and data entry applications. The various techniques covered under the general term character recognition [1] fall into either the on-line or off-line category, each having its own hardware and recognition algorithms.

Any Character Recognition system goes through numerous phases including: data acquisition, pre-processing, feature extraction, classification and post-processing where the most crucial aspect is the pre-processing which is necessary to modify the data either to correct deficiencies in the data acquisition process due to limitations of the capturing device sensor.

Data mining is an ambiguous term that has been used to refer to the process of finding interesting information in large repositories of data. "The science of extracting useful information from large data sets or databases" [2].Classification is one of the techniques which classify the given data based on many attribute given in the data base. This paper includes various clustering techniques such as K-means clustering, Hierarchical clustering and Density Based clustering are used for analyzing the datasets.

[II]. CLUSTERING

Cluster analysis is an iterated process of knowledge discovery and it is a multivariate statistical technique which identifies groupings of the data objects based on the inter-object similarities computed by a chosen distance metric .Clustering algorithms can be classified into two categories: Hierarchical clustering and Partitional clustering. The partitional clustering algorithms, which differ from the hierarchical clustering algorithms, are usually to create some sets of clusters at start and partition the data into similar groups after each iteration. Partitional clustering is more used than hierarchical clustering because the dataset can be divided into more than two subgroups in a single step but for hierarchy method, always merge or divide into 2 subgroups, and don't need to complete the dendrogram [3].

2.1 K-mean Clustering

The conventional K-mean algorithm is based on decomposition, most popular technique in data mining field. The concept of K-Means algorithm uses K as a parameter, Divide n object into K clusters, to create relatively high similarity in the cluster and, relatively low similarity between clusters. And minimize the total distance between the values in each cluster to the cluster center. The cluster center of each cluster is the mean value of the cluster. The calculation of similarity is done by mean value of the cluster objects. The measurement of the similarity for the algorithm selection is done by the reciprocal of Euclidean distance. That is to say, the closer the distance, the bigger the similarity of two objects, and vice versa.

Procedure of K-mean Algorithm

K-mean distributes all objects to K number of clusters at random;

- Calculate the mean value of each cluster, and use this mean value to represent the cluster;
- Re-distribute the objects to the closest cluster according to its distance to the cluster center;
- Update the mean value of the cluster, say, calculate the mean value of the objects in each cluster;
- Calculate the criterion function E, until the criterion function converges.

Usually, the K-mean algorithm criterion function adopts square error criterion, defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$$

In which, E is total square error of all the objects in the data cluster, p is given data object, m_i is mean value of cluster C_i (p and m are both multi-dimensional). The function of this criterion is to make the generated cluster be as compacted and independent as possible [4].

2.2 Hierarchical Clustering

Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The result of the algorithm is a tree of clusters, called dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level, a clustering of the data items into disjoint groups is obtained [5].

Agglomerative (bottom up)

1. Start with 1 point (singleton).
2. Recursively add two or more appropriate clusters.
3. Stop when k number of clusters is achieved.

Divisive (top down)

1. Start with a big cluster.
2. Recursively divides into smaller clusters.

3. Stop when k number of clusters is achieved.

General steps of Hierarchical Clustering*:* Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering (defined by S.C. Johnson in 1967) is this [6]:

- Start by assigning each item to a cluster, so that if we have N items, we now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
- Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now we have one cluster less.
- Compute distances (similarities) between the new cluster and each of the old clusters.
- Repeat steps 2 and 3 until all items are clustered into K number of clusters.

The merging criteria of clusters for hierarchical clustering are single link, average link and complete link use minimum, average and maximum distances between the members of two clusters, respectively [7].

2.3 Make Density Based Clustering

The cluster will be constructed based on the density properties of the database are derived from a human natural clustering approach. The clusters and consequently the classes are easily and readily identifiable because they have an increased density with respect to the points they possess. The elements of the database can be classified in two different types: the border points, the points located on the extremities of the cluster, and the core points, which are located on its inner region.

Let $x1, x2, \cdots xn$ be a sample of points generated by an underlying probability density function $P(x)$, which is assumed to be unknown. The cumulative distribution function is denoted as $P(x)$.

$$P(x) = \Sigma xi \leq xp(xi)$$

Let $p(x)$ denotes the estimate of $p(x)$ at x. We can estimate $p(x)$ by considering a window of width h centered at x. The width h is a parameter which denotes the spread or smoothness of the density estimate. If the spread is too large we get a more averaged value. If it is too small we do not have enough points in the window. For points within the window ($|z| \leq 1/2$) there is a net contribution of $1/hn$ to the probability estimate $p(x)$. On the other hand, points outside the window ($|z| > 1/2$) contribute 0.

$$K(z) = 1/\sqrt{2\pi} \exp\{-z2/2\}$$

Where z=x-xi/2

Here x (the center of the window) acts as the mean of the distribution, and h acts as the standard deviation of the distribution [8,9].

[III]. DATASET DESCRIPTION

The clustering algorithms are compared on Letter Image Recognition data and Multi-Feature Digit dataset. These datasets are available on UCI repository.

Letter Image Recognition dataset:

It is generated to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts), which were then scaled to fit into a range of integer values from 0 through 15 and 1 class attribute. The 16 integer attributes extracted from the raster scan images of the letters [10]. We tested our model with 530 instances of 5 letters (A, B, C, D, E, and F).

Multi-Feature Digit dataset:

The dataset has the features of handwritten numerals (‘0’—‘9’) extracted from a collection of dutchy utility maps. 200 patterns per class (for the total of 20000 patterns) have been digitized in binary images. These digits are represented in terms of 216 features using profile correlations and one class feature. This file contains 2000 instances in total. For test the model, the author

uses 494 instances of the Multi-Feature Digit Data [10].

[IV]. COMPARATIVE ANALYSIS

The K-means, Hierarchical and Make Density Based clustering are applied on Letter Image Recognition data and Multi-Feature Digit Data and their results are compared on the basis of accuracy and time complexity. The Table 1 shows the time taken by the clustering algorithms to make clusters when these datasets are deployed in the Weka Tool.

Table 1 Time taken by clustering algorithms to make cluster for given data sets

	Time for Letter Image Recognition Dataset (in sec)	**Time for Multi-Feature Digit Dataset (in sec)**
K-means	0.14	0.24
Hierarchical	0.99	1.92
Make Density Based Clustering	0.16	0.34

The time complexity for K-means, Hierarchical and Make Density Based clustering is shown in the Fig.1 .With the help of analysis, it is shown that K-means has required minimum time to make cluster for the both datasets in comparison with other clustering algorithms. K-means has maximum accuracy.

Table 2 Clustering algorithms result for Letter Image Recognition Dataset

Letter Image Recognition Dataset			
	Total number of Instances	**Correctly Classified Instances**	**Accuracy**
K-means	530	130	24.53
Hierarchical	530	117	22.08
Make Density Based Clustering	530	131	24.72

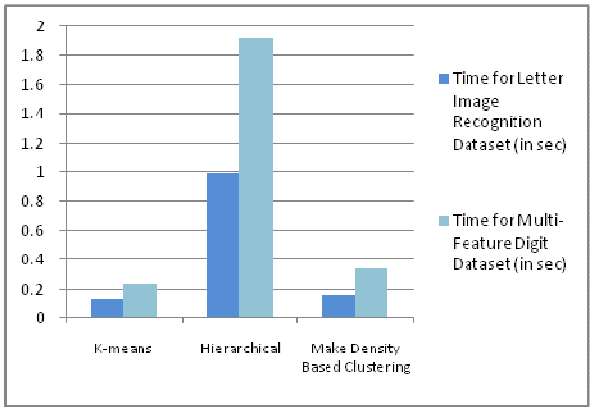

Figure 1 time take by the K-means, Hierarchical and Make Density Based clustering for datasets

First the datasets are deployed in Weka tool and then clustering algorithms are applied to the datasets with classes to cluster evaluation option. The accuracy of clustering algorithms in terms of correctly classified instances for Letter Image Recognition dataset and Multi-Feature Digit Dataset are shown in the Table 2 and Table 3 respectively. The accuracy of Letter Image recognition data and Multi-Feature digit data are compared for K-means, Hierarchical and Make Density Based clustering is shown graphically in Fig.2. An analysis of result shown in Fig. reveals that Make Density Based clustering has higher accuracy for both datasets. Whereas K-means has accuracy almost equal to Make Density Based clustering and Hierarchical clustering has lowest accuracy to recognizes the character.

Table 3 Clustering algorithms result for Multi-Feature Digit Dataset

Multi-Feature Digit Dataset			
	Total number of Instances	**Correctly Classified Instances**	**Accuracy**
K-means	494	229	46.36
Hierarchical	494	132	26.72
Make Density Based Clustering	494	230	46.56

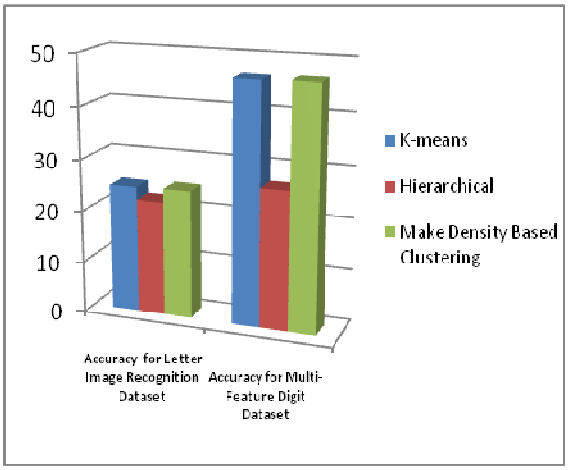

Figure 2 The cluster accuracy comparison of Letter Image Recognition data and Multi-Feature Digit Data for K-means, Hierarchical and Make Density Based Clustering.

[V]. CONCLUSION

The main objective of this paper is to make a comparative analysis of K-means, Hierarchical and Make Density Based clustering algorithms. It is important to remember that cluster analysis is an exploratory tool. While thousands of clustering algorithms are available and new ones continue to appear, we compare only three of them. K-means make clusters with minimum amount of time and good accuracy rate. Whereas Make Density Based Clustering recognize character with slightly higher accuracy than k-means but it takes more time to make clusters. In terms of time and accuracy K-means produces better results in comparison of all explained algorithms.

REFERENCES

[1] Arica N., Vural F.T.Y., "An Overview Of Character Recognition Focused On Off-line Handwriting" IEEE (1999), C99-06-C-203.

[2] Hand D., Mannila H., and Smyth P., "Principles of Data Mining", MIT Press, Cambridge, MA. (2001), ISBN 0-262- 08290-X.

[3] Gu J., Zhou J., Chen C., "An Enhancement of K-means Clustering Algorithm", in proceeding of Business Intelligence and Financial Engineering, 2009. BIFE '09, China.

[4] Wang J., Su X., "An improved K-means Clustering Algorithm", in proceeding of Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference,2011,China.

[5] Zhao Y., Karypis G., "Evaluation of hierarchical clustering algorithms for document datasets", the eleventh international conference on Information and knowledge management,2002,pp 515-524.

[6] Jain A.K., Murty M.N., Flynn P.J., "Data clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[7] Arai K., Barakbah A.R., "Hierarchical K-Means: An Algorithm For Centroids Initialization For K-Means", Reports of the Faculty of Science and Engineering, Saga University, (2007), Vol. 36, No.1, 25-31.

[8] Priyadarishini A., Karthik S., Anuradha J., B K Tripathy B.K., "Diagnosis of Psychopathology using Clustering and Rule Extraction using Rough Set", Advances in Applied Science Research, 2011, 2 (3), 346-362

[9] Margaret H. Dunham, "Data Mining- Introductory and Advanced Concepts", fourth edition, Pearson Education Publications, ISBN No-978-81-7758-785-2, 2009.

[10] archive.ics.uci.edu/ml/datasets.html.