Clustering Patient Records using Fuzzy C-Means and Hard C-Means Algorithm

Theint Theint Nwe Soe, Tin Tin Htwe
Computer University (Pathein)
theinttheintnwesoe88@gmail.com

Abstract

Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Most clustering algorithms, assign each data to exactly one cluster, thus forming a crisp (hard) partition of the given data, but fuzzy (soft) partition allows for degrees of membership, to which data belongs to different clusters. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of assigning these membership levels, and then using them to assign data elements to one or more cluster. This system is implemented clustering data by using Fuzzy C-Means (FCM) and Hard C-Means (HCM) clustering algorithms.

1. Introduction

Data clustering is considered an interesting approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups [1]. The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data into a small number of groups or categories in order to further facilitate its analysis. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Nevertheless, finding these groupings or trying to categorize the data is not a simple task for human unless the data is of low dimensionality (two or three dimensions at maximum). This is why some methods in soft computing have been proposed to solve this kind of problem. Those methods are called "Data Clustering Methods" and they are the subject of this paper.

Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols for example. Also if we can find groups of data, we can build a model of the problem based on those groupings. [2]

So, it is implemented in data clustering by using FCM and HCM algorithms to analyze the output result with validity and accuracy measure. The remainder of the paper is organized as follows: section 2 presents related work of this paper. Section 3 presents an overview of the system. Section 4 presents each of the two algorithms in detail. Validity index and accuracy measure of the clustering algorithms describe in Section 5. Section 6 includes system design. Section 7 is implementation of the system. Experimental results of the system presents in Section 8. A brief conclusion is presented at the last section.

2. Related Work

The system describes the current status and future prospects of applying data clustering approaches to data mining [1] and [2]. The authors in [3] and [4], describes about the k-means clustering methods same to HCM and algorithm explain clearly with step by step procedures. The system describes the FCM functional, detail of data clustering and FCM applied applications and FCM algorithm [5]. In [6], the authors discuss two facts of partitional clustering, namely on-line clustering technique and off-line clustering technique and describe briefly a few rule quality measures for both. In [7], the authors discuss comparative analysis of FCM and HCM algorithms based on complexities. The systems discuss some clustering algorithms and validity indices [8]. In [9], the system describes validity index suitable for FCM and HCM. In [11] and [12], the authors discuss briefly about thyroid gland.

3. System Overview

The system is presented with a training data set, which is used to find the cluster centers by analyzing all the input vectors in the training set. Once the cluster centers are found they are fixed, and they are used later to classify new input vector.

The common approach of all the clustering methods presented here is to find cluster centers that will represent each cluster. A cluster center is a way to tell where the heart of each cluster is located, so

that later when presented with an input vector, the system can tell which cluster this vector belongs to by measuring a similarity metric between the input vector and the cluster centers, and determining which cluster is the nearest or most similar one.

Some of the clustering techniques rely on knowing the number of clusters apriori. In that case the algorithm tries to partition the data into the given number of clusters. FCM and HCM clustering methods are of that type. If the number of clusters is not known, FCM and HCM clustering cannot be used.

Fuzzy C-Means clustering, which was proposed by Bezdek in 1973 [5] as an improvement over earlier HCM clustering. In this method, each data points belong to a cluster to a degree specified by a membership grade. This algorithm relies on finding clusters by trying to minimize a cost function of dissimilarity (or distance) measure.

Hard C-Means clustering, this has been applied to a variety of areas, including data compression, data preprocessing for system modeling and task decomposition in heterogeneous neural network architecture. In these methods, each data points are a member of one and only one cluster. As in FCM, HCM clustering relies on minimizing a cost function of dissimilarity measure.

4. Clustering Algorithms

4.1 FCM Algorithm

Fuzzy C-Means Clustering with the different that in FCM each data point belongs to a cluster to a degree of membership grade, while in HCM every data point either belongs to a certain cluster or not. So FCM employs fuzzy partitioning such that a given data point can belong to several groups with the degree of belongingness specified by membership grades that is to be minimized while trying to partition the data set.

The membership matrix U is allowed to have elements with values between 0 and 1. However, the summation of degrees of belongingness of a data point to all clusters is always equal to unity:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,...,n \tag{1}$$

The costs function for FCM:

$$J(U, c_1, c_2, ..., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{\ m} d_{ij}^{\ 2} \tag{2}$$

where,

u_{ij} = the degree of membership of x_i in the cluster j

d_{ij} = $\| c_i - x_j \|$ is the Euclidean distance between the i^{th} cluster center and the j^{th} data point

x_j = the j^{th} d-dimensional measured data

c_i = the d-dimension center of the cluster

m = fuzziness value (or) weighting exponent

The necessary conditions for Equation (2) to reach its minimum are

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{\ m} x_j}{\sum_{j=1}^{n} u_{ij}^{\ m}} \tag{3}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \tag{4}$$

The algorithm works iteratively through the preceding two conditions until the no more improvement is noticed. In a batch mode operation, FCM determines the cluster centers c_i and the membership matrix U using the following steps:

1. Randomly initializes the membership matrix (U) that has constraints in Equation (1).
2. Calculate centroids (c_i) by using Equation (3).
3. Compute dissimilarity between centroids and data points using Equation (2). Stop if its improvement over previous iteration is below a threshold.
4. Compute a new U using Equation (4). Go to Step 2.

The performance of the FCM algorithm depends on the initial membership matrix values: thereby it is advisable to run the algorithm for several times, each starting with different values of membership grades of data points.

4.2 HCM Algorithm

Hard C-Means clustering is an algorithm based on finding data clusters in a data set such that costs function (or an objection function) of dissimilarity (or distance) measure is minimized. In most cases, this dissimilarity measure is chosen as the Euclidean distance.

A set of n vectors x_j, j=1,…,n are to be partitioned into c groups G_i, i=1,…,c. The cost function, based on the Euclidean distance between a vector x_k in group j and the corresponding cluster center c_i, can be defined by:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \| x_k - c_i \|^2 \right) \quad (5)$$

The partitioned groups are defined by a cxn binary membership matrix U, where the element uij is 1 if the j^{th} data point x_j belongs to group i, and 0 otherwise. Once the cluster centers c_i are fixed, the minimizing u_{ij} for Equation (5) can be derived as follows:

$$u_{ij} = \begin{cases} 1 & \text{if } \| x_j - c_i \|^2 \leq \| x_j - c_k \|^2, \text{ for each } k \neq i, \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

which means that x_j belongs to group i if c_i is the closet center among all centers.

On the other hand, if the membership matrix is fixed, i.e. if u_{ij} is fixed, then the optimal center c_i that minimize Equation (5) is the mean of all vector in group i:

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \quad (7)$$

where $|G_i|$ is the size of G_i (or) $|G_i| = \sum_{j=1}^{n} u_{ij}$

The algorithm is presented with a data set x_i, i=1,…,n; it then determines the cluster centers c_i and the membership matrix U iteratively using the following steps:

(1) Initialize the centroids c_i, i=1,…,c. This is typically achieved by randomly selecting c points from among all of the data points.
(2) Determine the membership matrix U by Equation (6)
(3) Compute the dissimilarity function by using Equation (5). Stop if its improvement over previous iteration is below a threshold.
(4) Compute new centriods using by Equation (7). Go to Step 2.

The performance of the HCM algorithm depends on the initial positions of the cluster centers, thus it is advisable to run the algorithm several times, each with a different set of initial cluster centers.

5. Validity & Accuracy of Clustering Algorithms

Clustering validity is a concept to evaluate how good clustering results.

It is calculated using Davies-Bouldin Index.

5.1 Davies-Bouldin Index

DB index [9] is a function of the ratio of the sum of within-cluster scatter to between-cluster separation, it uses both the clusters and their sample means. The DB index is based on similarity measure of clusters (R_{ij}) whose bases are the dispersion measure of a cluster (s_i) and the cluster dissimilarity measure (d_{ij}). First, define the within i^{th} cluter scatter and the between i^{th} and j^{th} cluster as:

$$s_i = \frac{1}{\|c_{is}\|} \sum_{x \in c_i} d(x, v_i) \quad (8)$$

$$d_{ij} = d(v_i, v_j) \quad (9)$$

where,
d(x,y) = Distance between two data elements
c_i = i^{th} cluster
$\|c_i\|$ = Number of element in the i^{th} cluster
n_c = Number of clusters

Then the DB index is defined as

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \quad (10)$$

$$R_i = max_{j=1 \ldots n_c, i \neq j} (R_{ij}), i = 1 \ldots n_c \quad (11)$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (12)$$

DB index measures the average of similarity between each cluster and its most similar one. As the clusters have to be compact and separated the lower DB index means better cluster configuration. Minimize the DB index for achieving proper clustering.

5.2 Accuracy

Accuracy measure for output result is computed as follows:

$$r = \frac{\sum_{i=1}^{c} a_i}{n} \quad (13)$$

n=number of instance in the dataset

a_i=number of instance occurring true positive

In the formula a_i is the actual cluster result of machine learning based results and n is the manual classified data in the dataset.

6. System Design

In this system, number of clusters are predefined by 3 because of medical records set are cluster three groups by normal, hyperthyroidism function and hypothyroidism function.

For cluster process, user needs to choose the algorithm. And then the system can generate FCM algorithm and HCM algorithm based on user requirements training records set. And also the system evaluates the validity and accuracy of these two algorithms.

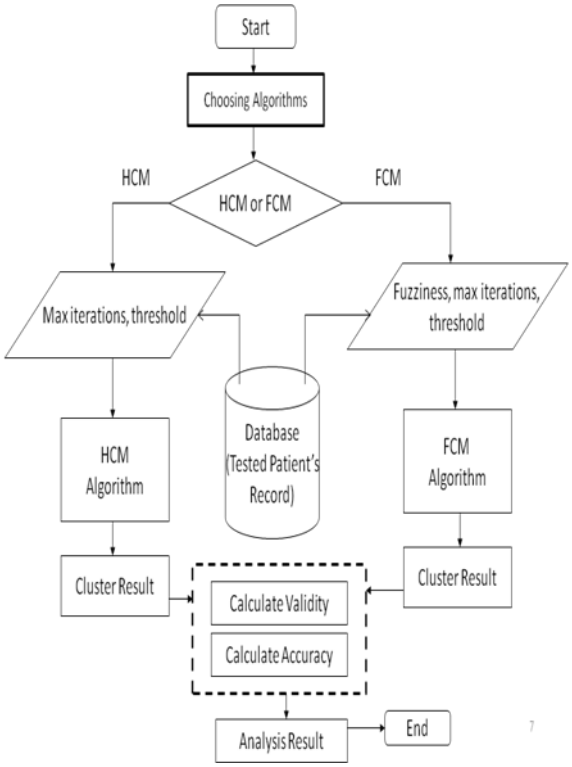

Figure 1. System Flow Diagram

In FCM algorithm portion, accept fuzziness (m), max-iteration (or) no: of records set, threshold (ϵ) from the user. Algorithm defines initialize membership matrix and calculates centroids. Then this algorithm calculates dissimilarity between centriods and input data and find the minimize value of dissimilarity. And it reassigns objects to clusters based on the distance between the objects and the centroids. The algorithm runs recursively until the terminating criteria is met. While running, the function prints a value that indicates the accuracy of the fuzzy clustering. When this value is less than the parameter threshold (ϵ), the function terminates. Otherwise, the algorithm recalculates the membership matrix.

In HCM algorithm portion, accept max-iterations (or) no: of records set and threshold (ϵ) form the user. The algorithm defines initialize clusters centroids and determines the membership matrix for data sets then computes the dissimilarity value between cluster center and data sample. Allocation of data points such that the distance is minimized. Then the algorithm runs recursively until any center no change. Otherwise the algorithm terminates.

For analysis process, the system also evaluate the validity and accuracy of clustering after clustered based on results set that is evaluated by using clustering algorithm. Finally, the system analyzes the performance of two algorithms.

7. Implementation of the System

The system clusters data with same group using the clustering algorithms and analyzes performance of two clustering algorithms Fuzzy C-Means (FCM) and Hard C-Means (HCM).

Table 1. Attributes of records set (Example- 10 records)

ID	TT3	TT4	T3	T4	TSH
1	107	10.1	2.2	0.9	2.7
2	113	9.9	3.1	2.0	5.9
3	127	12.9	2.4	1.4	0.6
4	109	5.3	1.6	1.4	1.5
5	105	7.3	1.5	1.5	-0.1
6	105	6.1	2.1	1.4	7.0
7	110	10.4	1.6	1.6	2.7
8	114	9.9	2.4	2.5	5.7
9	106	9.4	2.2	1.5	0.0
10	107	13.0	1.1	0.9	3.1

The system cluster medical record sets in this work. Medical record sets are the thyroid hormone test. There is TSH (Thyroid Stimulating Hormone), T3 (Triiodothyronine), T4 (Thyroxine), TT3 (Total Triiodothyronine) and TT4 (Total Thyroxine) in thyroid hormone. These are attributes in the dataset, moreover, patient's ID also include. The datasets are 215 records and 6 attributes. The system only uses one attributes is TSH when cluster the datasets into individual groups. Because of TSH is the most common reason for thyroid hormone tests [12]. Numbers of cluster are three groups: Normal, Hyperthyroidism and Hypothyroidism. If these groups how will cluster, the high TSH level indicates that the thyroid gland is filling so that the case of Hypothyroidism and the low level of TSH that is producing too much thyroid hormone called

Hyperthyroidism and then the TSH level does not above two conditions that is Normal condition. A standardized normal reference range for the TSH thyroid test is approximately 0.3 to 3.0. According the range, under range from 0.3 TSH test into Hyperthyroidism, above from 3.0 TSH test into Hypothyroidism and then if the TSH test is in this range into Normal. The above description is the portion of data clustering using clustering algorithms either FCM or HCM.

In this system, user can select the number of records for generating this system (at least 50 records). Although this system tests 215 records, it accepts maximum 1000 records set.

And, the system analyzes performance of clustering algorithms. Firstly, user chooses the algorithm. According the choosing algorithm, if the FCM is chosen, user must define the parameter value fuzziness (m →0.5 to 2.5), max-iterations or no: of records set (0 → 250, but must define into 215, otherwise, the system show the alarm if over 215), threshold (ϵ, between 0 and 1, but not use 1 because of FCM results and HCM results are same this threshold range). Then the system applies with clustering algorithms and produces the clustered results that are save in a file as FCM's results.

Next, another algorithm choose, if the user choose is HCM algorithm, user must define the parameter value max-iterations or no: of records set (0 → 250, but must define into 215, otherwise, the system show the alarm if over 215), threshold (ϵ, between 0 and 1). After apply, the system produces clustered results and saves it to a file as HCM's results.

After clustering, user evaluates the validity and the accuracy based on clustered results to compare two algorithms. The system runs on 10 times. The two algorithms run at 10 times with different value of parameter threshold. After run at once, user evaluates validity and accuracy based on clustered results set. After 10 times, the values of validity and accuracy of two algorithms analyze. Finally, analysis results show by graph.

8. Experimental Results

This system presents the study of partition clustering algorithms.

FCM starts with a random initialization of the partition matrix. It generally converges quite rapidly to accurate and robust in clustering by using the termination threshold value. When the FCM algorithm run, changes of threshold value effect directly on no: of iterations. The no: of iteration large if the threshold value reduces. It is noticed that very low values for threshold reduces the accuracy. Table 2 shows performance results of FCM algorithm.

Table 2. FCM clustering performance results

Threshold	No. of iteration	Validity	Accuracy (%)
0.0	174	1.48	70
0.1	293	3.59	69
0.2	366	3.31	66
0.3	384	1.86	67
0.4	521	3.05	69
0.5	683	0.72	66
0.6	588	0.49	67
0.7	658	2.03	65
0.8	735	0.70	65
0.9	625	3.26	65

HCM clustering starts with a random initialization of the cluster centers. So, several runs of the algorithms are advised to have better results. The changing of HCM results is a little then the results are stabled mostly. Although the values of threshold changed in the running of programs, accuracy results are not changed expect the result of threshold value 0.0. The results of no. of iterations same as accuracy results. Table 3 shows the performance results of HCM algorithm.

Table 3. HCM clustering performance results

Threshold	No. of iteration	Validity	Accuracy (%)
0.0	212	0.04	66
0.1	212	0.12	97
0.2	214	0.06	97
0.3	214	0.14	97
0.4	214	0.06	97
0.5	214	0.12	97
0.6	214	0.05	97
0.7	214	0.11	97
0.8	214	0.01	97
0.9	214	0.36	97

According the results of two algorithms, HCM algorithms suitable in this application because the algorithm give no change results and better accuracy and small amount of no: of iterations than FCM.

9. Conclusion

Two clustering algorithms have been reviewed in this paper, Fuzzy C-Means clustering and Hard C-

Means clustering. These approaches solve the problem of categorizing data by partitioning a data set into a number of clusters based on some similarity measure so that the similarity in each cluster is larger than among clusters. The two methods have been implemented and tested against a data set for medical diagnosis of thyroid disease. The comparative study done here is concerned with the validity and accuracy of each algorithm.

10. References

[1] Jang, J.-S.R., C.-T., Mizutani, E., Neuro-Fuzzy and Soft Computing-A Computational Approach to learning and Machine Intelligence, Prentice Hall

[2]A.K.Jain and R.C.Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.

[3]A.Likas,N.Vlassis, and J.Verbeek,"The global k-means clustering algorithm (Technical Report)",Computer Science Institute, University of Amsterdam, The Netherland. ISA-UVA-01-02-2001.

[4]J.A.Hartigan and M.A.Wong,"A k-means clustering algorithm", Applied Statistics, 1979, 28:100-108.

[5]J.C.Bezdek, R.Ehrlich, and W.Full, FCM: Fuzzy C-Means Algorithm, Computers and Geosciences, 1984

[6]K.M.Hammouda,"A comparative study of Data Clustering Techniques", Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2l3Gl.

[7]P.Kumar, P.Verma and R.Shrma,"Comparative analysis of fuzzy C Mean and Hard C Mean Algorithm", International Journal of information Technology and Knowledge Management, vol.2, 2010

[8]U.Maulik, and S.Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices", IEEE Trans, Pattern Analysis and Machine Intelligence, vol.24, no.12, 2002.

[9]D.L.Davies and D.W.Bouldin, "A cluster separation measure", IEEE Trans.PAMI, vol-1, 1979, pp.224-227

[10] Han J. and Kamber M., "Data Mining Concepts and Techniques". Morgan Kaufmann Publishers. 2001.

[11] S.Albayrak, and F.Amasyali, "Fuzzy C-Means Clustering on Medical Diagnostic Systems", Computer Engineering Department, Yidiz Technical University, 34349, Istanbul, Turkey.

[12] www.endocrineweb.com/thyroid.html, 2002.