FUZZY C-MEANS CLUSTERING ON MEDICAL DIAGNOSTIC SYSTEMS

Songül Albayrak[1] Fatih Amasyalı[2]

e-mail: songul@ce.yildiz.edu.tr e-mail: mfatih@ce.yildiz.edu.tr
Yıldız Technical University, Computer Engineering Department, 34349, Istanbul, Turkey

Key words: Unsupervised Clustering Methods, Fuzzy C-means, Hard K-means, Medical Diagnostic System

ABSTRACT

In this study, unsupervised clustering methods are examined to develop a medical diagnostic system and fuzzy c-means clustering is used to assign patients to the different clusters of thyroid diseases. The results are compared with the results of hard k-means clustering according to classification performance. This application shows that fuzzy clustering methods can be important supportive tool for the medical experts in diagnostic.

I. INTRODUCTION

According to rapid development on medical devices, the traditional manual data analysis has become inefficient and computer-based analyses are indispensable. Statistical methods, fuzzy logic, neural network and machine learning algorithms are being tested on many medical prediction problems to provide a decision support system.

Hard k-means algorithm executes a sharp classification, in which each object is either assigned to a class or not. The application of fuzzy sets in a classification function causes the class membership to become a relative one and an object can belong to several classes at the same time but with different degrees [3]. This is an important feature for medical diagnostic systems to increase the sensitivity.

In this work, unsupervised clustering methods were performed to cluster the patients into three clusters by using thyroid gland data obtained by Dr.Coomans [1]. To measure the thyroid gland functions, five different tests were applied to patients and the test results were used by other researchers for the classification purpose. In a very recent work, L. Ozyilmaz and T. Yildirim are investigated the supervised classification methods to develop a medical diagnostic system on this data [2]. In the work presented here, Fuzzy C-Means (FCM) and Hard C-Means (HCM) algorithms are used as an unsupervised clustering method to cluster the patients. As a result of clustering algorithms, patients' statuses are classified normal, hyperthyroid function and hypothyroid function.

II. HARD K-MEANS CLUSTERING

Hard k-means clustering, is also known as c-means clustering. The k-means algorithm partitions a collection of N vector into c groups (clusters G_i, i=1,..,c). The aim of that algorithm finding cluster centers(centroids) for each group. The algorithm minimizes a dissimilarity (or distance) function which is given in Equation 2.1.

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{k, x_k \in G_i} d(x_k - c_i) \qquad (2.1)$$

c_i is the centroid of cluster i;
$d(x_k - c_i)$ is the distance between i_{th} centroid(c_i) and k_{th} data point;

For simplicity, the Euclidian distance is used as dissimilarity measure and overall dissimilarity function is expressed as in Equation 2.2

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \| x_k - c_i \|^2 \right) \qquad (2.2)$$

Partitioned groups can be defined by an c*n **binary** membership matrix(U), where the element u_{ij} is 1 if the j_{th} data point x_j belongs to group i, and 0 otherwise. This explanation is formulated in Equation 2.3.

$$u_{ij} = \begin{cases} 1 & if \ \| x_j - c_i \|^2 \le \| x_j - c_k \|^2, \ for \ each \ k \neq i, \\ 0 & otherwise \end{cases} \qquad (2.3)$$

Since a data point can only be in a group, the membership matrix (U) has two properties which are given equation 2.4 and equation 2.5.

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, ..., n \qquad (2.4)$$

$$\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} = n \qquad (2.5)$$

Centroids are computed as the mean of all vectors in group i:

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \qquad (2.6)$$

$|G_i|$ is the size of G_i.

The k-means algorithm [4] determines the following steps with a data set x_j, j=1,..n;

Step 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.
Step 2. Determine the membership matrix U by Equation 2.3.
Step 3. Compute the dissimilarity function by using Equation 2.2. Stop if its improvement over previous iteration is below a threshold.

Step 4. Compute new centroids using by Equation 2.6. Go to step 2.

The performance of the algorithm depends on the initial positions of centroids. So the algorithm gives no guarantee for an optimum solution.

.

III. FUZZY C-MEANS CLUSTERING

Fuzzy C-means Clustering(FCM), is also known as Fuzzy ISODATA, is an clustering technique which is separated from hard k-means that employs hard partitioning. The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

FCM is an iterative algorithm. The aim of FCM is to find cluster centers (centroids) that minimize a dissimilarity function.

To accommodate the introduction of fuzzy partitioning, the membership matrix(U) is randomly initialized according to Equation 3.1

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,..., n \tag{3.1}$$

The dissimilarity function which is used in FCM is given Equation 3.2

$$J(U, c_1, c_2, ..., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} {u_{ij}}^m {d_{ij}}^2 \tag{3.2}$$

u_{ij} is between 0 and 1;
c_i is the centroid of cluster i;
d_{ij} is the Euclidian distance between i_{th} centroid(c_i) and j_{th} data point;
m ϵ [1,∞] is a weighting exponent.

To reach a minimum of dissimilarity function there are two conditions. These are given in Equation 3.3 and Equation 3.4.

$$c_i = \frac{\sum_{j=1}^{n} {u_{ij}}^m x_j}{\sum_{j=1}^{n} {u_{ij}}^m} \tag{3.3}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \tag{3.4}$$

Detailed algorithm of fuzzy c-means proposed by Bezdek in 1973[5]. This algorithm determines the following steps [4].

Step 1. Randomly initialize the membership matrix (U) that has constraints in Equation 3.1.
Step 2. Calculate centroids(c_i) by using Equation 3.3.
Step 3. Compute dissimilarity between centroids and data points using equation 3.2. Stop if its improvement over previous iteration is below a threshold.
Step 4. Compute a new U using Equation 3.4. Go to Step 2.

By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the "right" location within a data set.

FCM does not ensure that it converges to an optimal solution. Because of cluster centers (centroids) are initialize using U that randomly initialized.(Equation 3.3).

Performance depends on initial centroids. For a robust approach there are two ways which is described below.
1-) Using an algorithm to determine all of the centroids. (for example: arithmetic means of all data points)
2-) Run FCM several times each starting with different initial centroids.

We preferred the first one and obtained the better performance on thyroid gland data.

IV. THE FUNCTIONS AND PROPORTIES OF THE THYROID GLAND

The thyroid gland is the biggest gland in the neck. It is situated in the front neck bellow the skin and muscle layers. The thyroid gland takes the shape of a butterfly with the two wings being represented by the left and right thyroid lobed which wrap around the trachea. The sole function of the thyroid is to make thyroid hormone. This hormone has an effect on nearly all tissues of the body where it increases cellular activity. The function of the thyroid therefore is to regulate the body's metabolism [6].

The thyroid gland is prone to several very distinct problems, some of which are extremely common. Production of too little thyroid hormone causes hypo-thyroidism or production of too much thyroid hormone causes hyper-thyroidism.

In this work, thyroid database [7] is investigated to cluster by unsupervised methods. This data set contains 3 classes and 215 samples. These classes are assigned to the values that correspond to the hyper-, hypo- and normal function of the thyroid gland. Class distribution or number of instance for normal class is 150, for hyper class is 35 and for hypo class is 30. The followings give the 5 tests which are applied to patients to measure the thyroid function. 5 dimensional feature vector is obtained as $x=[x_1,x_2,x_3,x_4,x_5]$ from the applied tests

1-T3-resin uptake test (A percentage)
2-Total Serum thyroxin as measured by the isotopic displacement method.
3-Total Serum triiodothyronine as measured by radioimmuno assay
4-Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay
5-Maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value.

V. EXPERIMENTAL RESULTS

Fuzzy c-means and hard c-means algorithms are used to assign the patients to different clusters of thyroid diseases. Hard c-means algorithm is applied to thyroid gland data and 168 correct classified samples are obtained out of 215 samples. Fuzzy c-means give better results with 180 correct classified samples. Table 1 gives the comparison of two classifiers.

Table 1. Comparison of Classifiers

Classifier	# of Correct Classified Samples	Correct Classification Rate
Hard C-Means	168	78.1%
Fuzzy C-Means	180	83.7%

In Figure 1, the thyroid gland data is put in order according to their membership degrees for graphical interpretation. In the graphical presentation, hypo and hyper thyroid patients are composed into one cluster.

The application of fuzzy sets in a classification function causes the class membership to become a relative one and an object can belong to several classes at the same time but with different degrees. This is an important feature for medical diagnostic systems to increase the sensitivity.

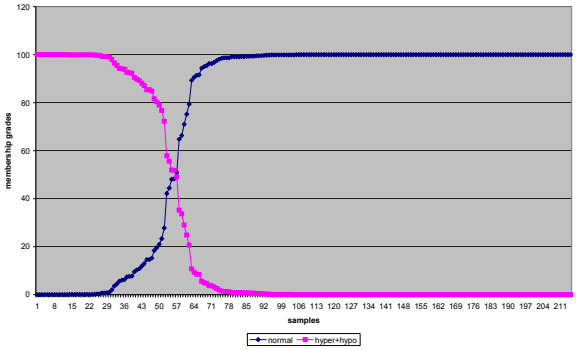

Figure 1. Graphical presentation of Fuzzy C-means clustering.

For the membership degrees close to 0.5 are the suspicious cases (shaded area in Table 2) to assign the sample to one cluster. Therefore fuzzy c-means clustering for medical diagnostic systems is more reliable than the hard one.

Table 2. Membership degrees of each samples to normal, hypo and hyper clusters normalized with 100

Samples	1	2, 3....	51	52	53	54	55	56	57	58	59	.....	215
μ_1(normal)	**0**	.....	**19,6**	**21**	**23,3**	**27,8**	**44,4**	**48,1**	**50,9**	**66,3**	**71**	.....	**100**
μ_2(hypo)	**100**	.....	**0**	**79**	**0**	**0,1**	**0**	**0**	**49**	**0**	**0**	.....	**0**
μ_3(hyper	**0**	.....	**80,3**	**0**	**76,7**	**72,2**	**55,6**	**51,9**	**0**	**33,7**	**29**	.....	**0**

Suspicious samples

VI. CONCLUSION

In this study, we use fuzzy c-means and hard k-means algorithms to cluster the thyroid gland data. In medical diagnostic systems, fuzzy c-means algorithm gives the better results than hard-k-means algorithm according to our application. Another important feature of fuzzy c-means algorithm is membership function and an object can belong to several classes at the same time but with different degrees. This is a useful feature for a medical diagnostic system. At a result, fuzzy clustering methods can be important supportive tool for the medical experts in diagnostic.

REFERENCES

1. Coomans, I. Broeckaert, M. Jonckheer, D.L. Massart: Comparison of Multivariate Discrimination Techniques for Clinical Data - Application to the Thyroid Functional State. Methods of Information in Medicine, Vol.22, (1983) 93- 101
2. L. Ozyilmaz, T. Yildirim, Diagnosis of Thyroid Disease using Artificial Neural Network Methods, Proceedings of the 9'th International Conference on Neural Information Processing (ICONIP'02) (2002)
3. G. Berks, D.G. Keyserlingk, J. Jantzen, M. Dotoli, H. Axer, Fuzzy Clustering- A Versatile Mean to Explore Medical Database, ESIT2000, Aachen, Germany
4. J.-S. R. Jang, C.-T. Sun, E.Mizutani, Neuro-Fuzzy and Soft Computing, p (426-427)Prentice Hall, 1997
5. Jim C. Bezdek. Fuzzy Mathemathics in Pattern Classification. PhD Thesis, Applied Math. Center, Cornell University, Ithaca, 1973.
6. www.endocrineweb.com/thyroid.html, 2002
7. www.ics.uci.edu/pub/ml-repos/machine-learning-database/, 2001