International Electronic Journal of Pure and Applied Mathematics

Volume 11 No. 1 2017, 51-58
ISSN: 1314-0744
url: http://www.e.ijpam.eu
doi: http://dx.doi.org/10.12732/iejpam.v11i1.5

CLASSIFICATION OF THYROID SAMPLES BY FUZZY CLUSTERING ALGORITHMS

B. Venkataramana[1], L. Padmasree[2],
M. Srinivasa Rao[3], B.N.V. Satish[4], G. Ganesan[5] §

[1]Department of Computer Science & Engineering
Holy Mary Institute of Technology
Bogaram, Telengana, INDIA

[2]Department of Electronics & Communications Engineering
VNR Vignana Jyothi Institute of Engineering & Technology
Bachupalli, Telengana, INDIA

[3]School of Information Technology
Jawaharlal Nehru Technological University
Hyderabad, Telangana, INDIA

[4]Department of Humanities & Basic Sciences
Aditya Engineering College
Surampalem, Andhra Pradesh, INDIA

[5]Department of Mathematics
Adikavi Nannaya University
Rajahmundry, Andhra Pradesh, INDIA

Abstract: Due to fast growth in technology, conventional classification methods are difficult to analyze accurate diagnosis without ambiguities. Since the states are vague in medicine the fuzzy methods are supportive rather than crisp ones. The objective of this paper is to analyze the classification performance in Thyroid Samples in terms of percentage of correctness among the non-fuzzy and fuzzy clustering algorithms k-means and fuzzy c-means respectively.

AMS Subject Classification: 03B52, 68W25
Key Words: fuzzy c-means, k-means, fuzzy clustering

Received: October 10, 2016

§Correspondence author

International Electronic Journal of Pure and Applied Mathematics – IEJPAM, Volume 11, No. 1 (2017)

1. Introduction

Clustering is one of the important phenomenons in soft computing which creates clusters of most identical featured objects in a group of data. A cluster of objects can be treated collectively as one group and so may be considered as form of data classification. Clustering data streams attracted many researches since the applications that generate data streams have become more popular. Clustering is also often called as Classification. Clustering is an important tool in data analysis, image processing, data mining, pattern recognition, medical diagnostics and etc [1].

Thyroid gland is one of the largest of endocrine gland, weighing 15-20 g in adults. Thyroid secretes two major hormones thyroxine and tri-idothyronine, commonly called t4 and t3 respectively. Thyroid secretion is controlled primarily by thyroid stimulating hormone [TSH] secreted by pituitary gland. Thyroid gland also secretes calcitonin a hormone involved in calcium metabolism. It consists of large number of closed follicles that are filled with a secretory substance called colloid and lined by cuboidal epithelium. Thyroid gland which is in a butterfly-shape is one of the largest endocrine glands located in the lower front of the neck [2]. This gland produces thyroid hormones to regulate the bodys metabolism and calcium balance. It helps to maintain the working conditions of brain, heart, muscles and other organs and helps body to stay warm and use energy. A symptom is a physical or laboratory finding that indicates the presence of a disease and hence can be considered as an aid in diagnosis. A cluster diagnosis one of the main tasks is grouping the symptoms to one syndrome. In this regard clustering analysis is well known as an effective and efficient tool in medicine.

Due to the rapid growth in technology, conventional classification methods are quite difficult to analyze accurate diagnosis without ambiguities. Since the conditions are vague in medicine, fuzzy methods are supportive rather than crisp.

The fuzzy cluster analysis is an iterative method. In this method memberships are assigned to the objects ranged between 0 and 1 by means of a membership function. This feature becomes a relative one and simultaneously more than one class or cluster can have the same object but with different degrees. These algorithms look for the cluster prototypes by optimizing the objective function (a function which is used to find the distance between the prototype and the object).

In this paper, the authors aim to provide the clustering results of the Thyroid diseases clusters by implementing non-fuzzy and fuzzy clustering methods namely k-means, Fuzzy c-Means (FCM) methods respectively.

International Electronic Journal of Pure and Applied Mathematics – IEJPAM, Volume 11, No. 1 (2017)

2. Materials and Methods

2.1. The Dataset

The Thyroid gland dataset was gathered from the UCI Machine Learning Repository [3]. The dataset contains 215 samples with 5 attributes or lab measurements each. The samples are classified into three different classes according to the Thyroid functions: Normal (150 samples), Hyperthyroid (35 samples) and Hypothyroid (30 samples). The 5 attributes are the lab tests to measure the thyroid function. These attributes are T3-resin uptake test (A percentage), value of total serum thyroxin given by the isotopic displacement method, total serum triiodothyronine value given by radioimmunoassay, value of basal thyroid stimulating hormone (TSH) given by radioimmunoassay and after injection of 200 micro grams of thyrotrophic-releasing hormone the maximal absolute difference of TSH value as compared to the basal value.

2.2. K-means Algorithm

The K-Means [4] is one of the famous hard clustering algorithm. It takes the input parameter k, the number of clusters, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to recalculate k new centroids. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \parallel x_i^{(j)} - c_j \parallel^2,$$

where $\parallel x_i^{(j)} - c_j \parallel^2$ is a choosen distance measure between a data point $x_i^{(j)}$ and the cluster center, c_j is an indicator of the distance of an n data points from their respective cluster centers.

The algorithm is:

International Electronic Journal of Pure and Applied Mathematics – IEJPAM, Volume 11, No. 1 (2017)

Step 1: Select K points as initial centroids.

Step 2: Repeat.

Step 3: Form k clusters by assigning all points to the closest centroid.

Step 4: Re-compute the centroid of each cluster.

Step 5: Until the centroids do not change.

K-means algorithm is significantly sensitive to the initial randomly selected cluster centers. The algorithm can be run multiple times to reduce this effect. The K-Means is a simple algorithm that has been adapted to many problem domains and it is a good candidate to work for a randomly generated data Repeat 2 and 3 until no change in each cluster center.

2.3. Fuzzy c-Mean Clustering

Fuzzy c-Means [4] algorithm is one of the most popular clustering fuzzy clustering method. Consider a dataset Z with N observations is an n-dimensional row vector. $z_k = [z_{k_1}, z_{k_2}, ..., z_{k_n}] \in \Re^n$. The dataset Z is represented as N x n matrices. In medical diagnosis the rows of Z represents patients and the columns are symptoms or laboratory measurements for these patients. The partition of the dataset Z into c ($1 \leqslant c \leqslant N$) clusters is represented by the fuzzy partition matrix $U = [\mu_{ik}]_{cxN}$. The fuzzy partitioning space for Z is the set

$$M_{fc} = \left\{ U \in \Re^{cxN} / \mu_{ik} \in [0,1],\ \forall i,k;\ \sum_{i=1}^{c} \mu_{ik} = 1, \forall k;\ 0 < \sum_{k=1}^{N} \mu_{ik},\ \forall i \right\}. \quad (1)$$

Fuzzy c-Mean model achieves its partitioning by the iterative optimization of its objective function given as

$$\underbrace{\min}_{U,V} \left\{ J(Z;U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m \parallel z_k - v_i \parallel_A^2 \right\}, \quad (2)$$

where $U = [\mu_{ik}] \in M_k$.

Here $m \in [1, \infty)$ is a parameter that determines the degree of fuzziness, $V = [v_1, v_2, ..., v_c]$ where $v_i \in \Re^n$ is a vector of (unknown) cluster prototypes (centers). The prototypes, the membership functions and the euclidean distance metric are calculated by the equations (3), (4), (5) respectively.

$$v_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m z_k}{\sum_{k=1}^{N} (\mu_{ik})^m}, \quad 1 \leq i \leq c, \quad (3)$$

International Electronic Journal of Pure and Applied Mathematics – IEJPAM, Volume 11, No. 1 (2017)

$$\mu_{ik} = (\sum_{j=1}^{c} (\frac{D_{ikA}}{D_{jkA}})^{\frac{2}{m-1}})^{-1}, \quad 1 \leq i \leq c,\ 1 \leq k \leq N, \tag{4}$$

$$D_{ikA}^2 = \| z_k - v_i \|_A^2 = (z_k - v_i)^T A (z_k - v_i),\ 1 \leq i \leq c,\ 1 \leq k \leq N. \tag{5}$$

When the objective function converges to a local minimum the iteration terminates. Detailed algorithm was proposed [5] is given below.

The algorithm is given by the following basic steps.

Step 1: Randomly initialize partition matrix U, number of clusters c, weighting parameter m and the termination tolerance $\varepsilon > 0$.

Step 2: Determine the fuzzy cluster prototypes by using the equation (3).

Step 3: update the membership matrix by using the equation (4).

Step 4: Compare the membership matrices of previous and after the iteration and repeat from step 2 until it meets the convergence criteria.

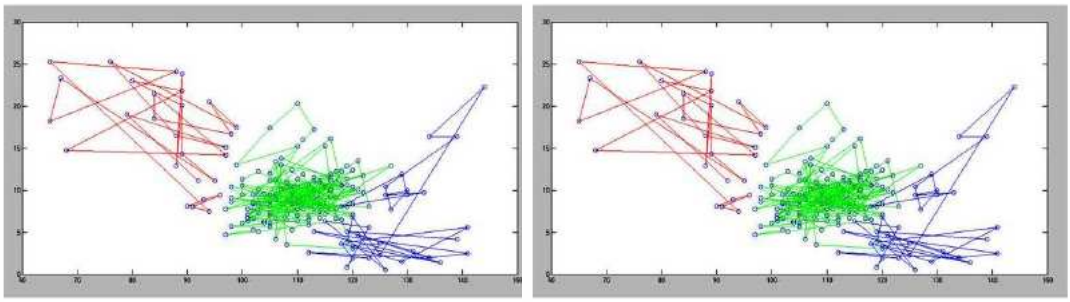
In fuzzy clustering, FCM is a popular clustering method but it has also some drawbacks. For example, if the method is used to partition two clusters and there is an object which is equidistance from two centers then according to the constraint on the membership value it assigns equal membership value regardless of the actual belonging to a cluster. These points are called as noise points.

3. Results and Discussion

The algorithms were implemented in MATLAB version R2012a. The data set contains 215 samples classified as three different types of thyroid functions. Each sample consists of 5 different real valued continuous lab measurements. These measurements are T3-resin uptake test (A percentage), value of total serum thyroxin given by the isotopic displacement method, total serum tri-iodothyronine value given by radioimmunoassay, value of basal thyroid-stimulating hormone (TSH) given by radioimmunoassay and after injection of 200 micro grams of thyrotropin-releasing hormone the maximal absolute difference of TSH value as compared to the basal value.

All samples are labeled by numbers 1 to 215. The samples from 1 to 150 are classified as Normal, 151 to 185 are classified as Hyperthyroid and 186 to 215 are classified as Hypothyroid. The algorithms K-means, FCM generates 3 clusters. These algorithms run several times in order to achieve good results. FCM generates three clusters corresponding to Normal, Hyperthyroid and Hypothyroid containing 128, 39 and 48 samples respectively. The cluster which is associated with Normal contains 7 samples that belong to Hyperthyroid and 3 samples that belong to Hypothyroid clusters are wrongly grouped. The cluster which is associated hyperthyroid contains

International Electronic Journal of Pure and Applied Mathematics – IEJPAM, Volume 11, No. 1 (2017)



(a) k-mean cluster (b) FCM cluster

Figure 1: Clustering results

Table 1: The clustering results obtained by K-means

Clusters	k-means clustering			
	correct	incorrect	total	percentage of correctness
Normal	137	15	152	91.33
Hyperthoroid	23	6	29	65.71
Hypothyroid	24	10	34	80.00

13 samples that belong to Normal and 1 sample that belongs to Hypothyroid clusters are wrongly assigned with Hyperthyroid, and 19 samples that belong to Normal and 3 samples that belong to Hyperthyroid clusters are wrongly classified in to the cluster associated with Hypothyroid.

K-means generates three clusters corresponding to Normal, Hyperthyroid and Hypothyroid containing 152, 29 and 34 samples respectively. The cluster which is associated with Normal contains 9 samples that belong to Hyperthyroid and 6 samples that belong to hypothyroid clusters are wrongly grouped. The cluster which is associated hyperthyroid contains 6samples that belong to Normal clusters are wrongly assigned with Hyperthyroid, and 7 samples that belong to Normal and 3 samples that belong to Hyperthyroid clusters are wrongly classified in to the cluster associated with Hypothyroid.

The clustering results of the two fuzzy methods are given in table 1, table 2 and the graphs are shown in Figures 1(a) and 1(b).

In the Figure 1(a) shows the results of K-means clustering algorithm applied to thyroid data. In this Green colored line indicated normal, Red line indicated hyperthyroid and blue line indicated hypothyroid.

In the Figure 1(b) green line connects hyperthyroid, red line connects normal and blue line connects hypothyroid samples.

International Electronic Journal of Pure and Applied Mathematics – IEJPAM, Volume 11, No. 1 (2017)

Table 2: The clustering results obtained by Fuzzy C Means

Clusters	Fuzzy c-mean clustering			
	correct	incorrect	total	percentage of correctness
Normal	118	10	128	78.67
Hyperthoroid	25	14	39	71.43
Hypothyroid	26	22	48	86.16

4. Conclusion

In this work, authors aim to examine the classification production of various fuzzy clustering methods in medical diagnosis. The authors implemented the fuzzy clustering algorithms, Fuzzy c-Means (FCM), and k-means algorithms and discussed the results Fuzzy c-means model behaves better than k-means model. The two algorithms performed well and the performance and correctness were comparable. The correctness percentage of Hyperthyroid and Hypothyroid clusters generation were equal in all the models. The k-means generates better cluster associated to Normal by acquiring high percentage of correctness than the FCM model. In overall classification performance, FCM placed high among the k-means clustering algorithm. As a result, fuzzy clustering algorithms can become an important tool in medical diagnosis. Further research is needed on improving the performance in classification and correctness in order to achieve at most satisfaction.

References

[1] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[2] www.endocrineweb.com/thyroid.html

[3] I. Coomans, M. Broeckaert, D. Jonckheer, L. Massart, Comparison of multivariate discrimination techniques for clinical data - Application to the thyroid functional state, *Methods of Information in Medicine*, **22** (1983), 93-101.

[4] R/ Jipkate Bharati, V.V. Gohokar, A comparative analysis of fuzzy c-means clustering and k means clustering algorithms, *Int. J. Comput. Eng. Res.*, **2** (2012), 737-739.

[5] James C. Bezdek, Robert Ehrlich, William Full, FCM: The fuzzy c-means clustering algorithm, *Computers & Geosciences*, **10**, No-s: 2-3 (1984), 191-203.

[6] B. Simhachalam, G. Ganesan, Performance comparison of fuzzy and non-fuzzy classification methods, *Egyptian Informatics Journal*, **17**, No. 2 (2016), 183-188.

[7] B. Simhachalam, G. Ganesan, Fuzzy clustering algorithms in medical diagnostics, *Wulfenia Journal*, **22**, No. 7 (2015), 308-316.

International Electronic Journal of Pure and Applied Mathematics – IEJPAM, Volume 11, No. 1 (2017)