Epileptic Seizure Recognition by Classification

Shazana Affandi, Athirah Zaini, Fatihah Zamri, Asiyah Zukfily

Faculty of Information Communication and Technology, UTeM

***Abstract* – Fuzzy clustering is a clustering method in data mining which group most similar data in a dataset. In this project, the comparison between Fuzzy C-Means (FCM) clustering algorithm and Hard C-Means (HCM) algorithm has been done to check for the most suitable algorithm for the dataset. Based on the Epileptic Seizure Recognition dataset, there are 4097 electroencephalograms (EEG) readings per patient over 23.5 seconds, with 500 patients in total before divided equally into 23 chunks per patient. With all of these, the fuzziness of the data is figured out by k-Mean clustering and then feature selection is done to find the most suitable attribute among the data in the dataset. As the result there are 20 attributes that are the best of all. After that, FCM and HCM is performed with the 20 attributes and 150 examples taken from the dataset. After all the process is done, it has been concluded that HCM clustering is more suitable than the FCM for this dataset to determine the fuzziness with its graphs show that HCM has cluster plot with bigger in distance better than FCM. The performance has been evaluated using R-Studio.**

I. INTRODUCTION

Epilepsy is a common brain disorder according to an estimate of the World Health Organization that affects almost 60 million people around the world. The seizures are causes from a transient and unexpected electrical disturbance of the brain and excessive neuronal discharge that is obvious in the electroencephalogram (EEG) signal representative of the electrical activity of the brain. To check the abnormality in brain activity, the EEG signal is classified by using some classifiers.

In this paper, two clustering algorithms are used to cluster the epilepsy dataset which is Fuzzy c means (FCM) and Hard c means (HCM). Fuzzy C Means (FCM) is a technique of clustering which allows one piece of data to belong to two or more clusters meanwhile Hard C Means (HCM) is non-fuzzy or hard clustering. HCM also known as k-means.

So, in this paper we discussed how to conduct the clustering for the Epileptic Seizure Recognition and the rest of paper is organized as follows: Section II contains the dataset description. Section III contains FCM and HCM algorithm. Section IV contains implementation, testing and results and section V concludes our work.

II. DATASET DESCRIPTION

The dataset used for this project is Epileptic Seizure Recognition from Kaggle.com. This dataset includes 4097 electroencephalograms (EEG) readings per patient over 23.5 seconds, with 500 patients in total. The 4097 data points were then divided equally into 23 chunks per patient; each chunk is translated into one row in the dataset. Each row contains 178 readings that are turned into columns; in other words, there are 178 columns that make up one second of EEG readings. All in all, there are 11,500 rows and 178 columns with the first being patient ID and the last column containing the status of the patient, whether the patient is having a seizure or not.

Using X and Y as the variables, there are 178 input in total which consists of X1 until X178 while Y as the output contains 5 stage of results.

y	Status
1	Recording of seizure activity
2	They recorder the EEG from the area where the tumor was located

3	Yes they identify where the region of the tumor was in the brain and recording the EEG activity from the healthy brain area
4	eyes closed, means when they were recording the EEG signal the patient had their eyes closed
5	eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open

All subjects falling in classes 2, 3, 4, and 5 are subjects who did not have epileptic seizure. Only subjects in class 1 have epileptic seizure.

III.FCM and HCM ALGORITHM

A. Fuzzy C-Means

Fuzzy C-Means (FCM) is a technique of clustering which allows one piece of data to belong to two or more clusters. FCM uses concepts from field of fuzzy logic and fuzzy set theory. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. Below are the steps of FCM algorithm.

1. Fix the value of c where (2<=c<n) and select value for parameter m'. Initialize the partition matrix $U^{(0)}$. Each step will be labeled as r, where r=0, 1, 2.

2. Calculate the c center vector {vij} for each step

$$\mathrm{V_{ij}} = \frac{\sum_{k=1}^{n} u_{ik}^{m'} \times x_{kj}}{\sum_{k=1}^{n} u_{ik}^{m'}}$$

3. Calculate the distance matrix $D_{[c,n]}$.

$$\mathrm{D_{ij}} = [\sum_{j=1}^{m} x_{kj} - v_{ij}^{2}\,]^{1/2}$$

4. Update the partition matrix for the r^{th} step, $U^{®}$ as follow:

$$u_{ik}^{r-1} = \frac{1}{\sum_{j=1}^{c} [\frac{d_{ik}^{r}}{d_{jk}^{r}}]^{\frac{2}{[m-1]}}}$$

If $\left|\left|U^{(k+1)} - U^{(k)}\right|\right| < \delta$ then stop, otherwise return to step 2 by iteratively updating the cluster centers and the memberships grades for data point.

B. Hard C-Means

Hard C-Means (HCM) is non-fuzzy or hard clustering. HCM also known as k-means. Partial membership is not allowed in this clustering technique. HCM is used to classify data in a crisp sense where every data point will be assigned to one and only one data cluster. The sum of membership grades of each data point in all clusters is equal to one and membership grade of a specific data point in a specific cluster is one and in all the remaining clusters its membership grade is zero. Below are the steps of FCM algorithm.

1. Fix the value of c where (2<=c<n) and initialize the U matrix.
   r=0, 1, 2, 3…
   $$U^{(0)} \in M_c$$
2. Calculate the center vectors { $V^{®}$ with $U^{®}$ }

3. Update $U^{®}$ and then calculate the updated characteristics function for all i, k.

   $X_{ik}\ (r+1) = 1, d_i(r) = \min d_{jk}\ (r)$ for all j $\in$ c0, otherwise

4. If $\left|\left| U^{(r-1)} - U^{®} \right|\right| <= \delta$ (tolerance level)

Stop otherwise, set r=r+1 and return to step 2. The notation || || is any matrix norm such as the Euclidean norm.

IV.IMPLEMENTATION, TESTING AND RESULTS.

A. Implementation

This phase will focus on how to implement the coding for the visualize the fuzziness of the data, HCM and FCM of the project. RStudio was used in the making of the three main implementation. In order to run the coding, the file was saved as R format.

B. Testing

1. Visualize Fuzziness of Dataset

   This step is to test the fuzziness of the dataset. It is because we want to know the quality of the dataset weather it suitable for HCM or HCM.

```
> library("FactoMineR")
> library("VIM")
> library("factoextra")
> library("tidyverse")
> library("cluster")
> library(magrittr)
> library(dplyr)
> Epileptic <- read.csv("C:/Users/shaza/Desktop/Epileptic.csv")
> ep.kmns1 <- kmeans(newEpileptic, centers = 3, nstart = 25)
> fviz_cluster(ep.fanny, data = newEpileptic)
> ep.sil1 <- silhouette(ep.kmns1$cluster, dist(Epileptic))
> fviz_silhouette(ep.sil1)
```

Figure 4.1 Coding for Fuzziness Visualization

2. Feature Selection

   Our dataset contains 178 attributes with 1150 data. The dataset provided 5 levels of people weather they have epileptic seizure or not. In order to do HCM and FCM, we used feature selection to choose which attributes is more important and the dataset is separated into 3 levels which each of every level contained 50 data. The total of overall data is 150.

```
library(Boruta)
library(caret)

> Epileptic <- read.csv("C:/Users/shaza/Desktop/Epileptic.csv"
>   View(Epileptic)
> rPartMod <- train(y ~ ., data = Epileptic, method = "rpart")
> View(rPartMod)
> rpartImp <- varImp(rPartMod)
> print(rpartImp)
```

Figure 4.2 Coding of Feature Selection

3. Fuzzy C-Means

   In order to find FCM, we only used chosen attributes from features selection.

```
> library("FactoMineR")
> library("factoextra")
> library("tidyverse")
> library("cluster")
> library(magrittr)
> Epileptic <- read.csv("C:/Users/shaza/Desktop/Epileptic.csv")
>   View(Epileptic)
> newEpileptic <- Epileptic %>% select(X121, X153, X92, X91, X120, X103, X24, X113, X44, X130, X78, X102, X133, X165, X73, X146, X124, X129, X84, X4)
> ep.fanny <- fanny(newEpileptic, 3)
> fviz_cluster(ep.fanny, data = newEpileptic)
> fviz_silhouette(ep.fanny, data = newEpileptic)
```

Figure 4.3 Coding for Fuzzy C-Means

4. Hard C-Means

   In order to find HCM, we only used chosen attributes from features selection.

```
> library("FactoMineR")
> library("VIM")
> library("factoextra")
> library("tidyverse")
> library("cluster")
> library(magrittr)
> library(dplyr)
> Epileptic <- read.csv("C:/Users/shaza/Desktop/Epileptic.csv")
> newEpileptic <- Epileptic %>% select(X121, X153, X92, X91, X120, X103, X24, X113, X44, X130, X78, X102, X133, X165, X73, X146, X124, X129, X84, X4)
> ep.kmns <- kmeans(newEpileptic, centers = 3, nstart = 25)
> str(ep.kmns)
> fviz_cluster(ep.kmns, data = newEpileptic)
> ep.sil <- silhouette(ep.kmns$cluster, dist(newEpileptic))
> fviz_silhouette(ep.sil)
cluster size ave.sil.width
1       1   10         -0.12
2       2    3          0.13
3       3  137          0.66
```

Figure 4.4 Coding for Hard C-Means

C. Results

The results shown below:

1. Visualize Fuzziness of Dataset

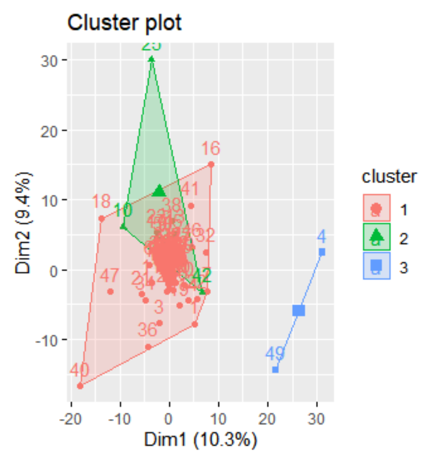

Figure 4.5 Cluster Plot for Fuzziness Visualization

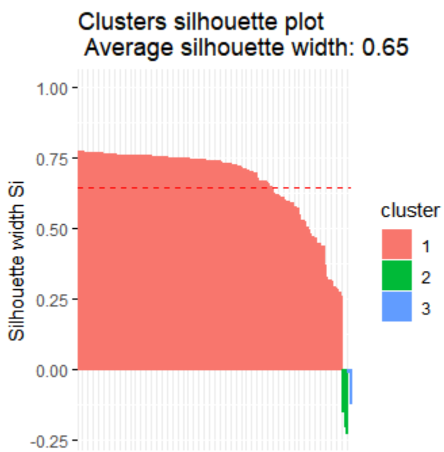

Figure 4.6 Silhouette of Fuzziness Visualization

2. Features Selection

```
rpart variable importance

only 20 most important variables shown (out of 178)

Overall
X121  100.00
X153   96.68
X92    48.73
X91    48.15
X120   47.36
X103   46.20
X24    45.72
X113   45.71
X44     0.00
X130    0.00
X78     0.00
X102    0.00
X133    0.00
X165    0.00
X73     0.00
X146    0.00
X124    0.00
X129    0.00
X84     0.00
X4      0.00
```

Figure 4.8 Results of Feature Selection

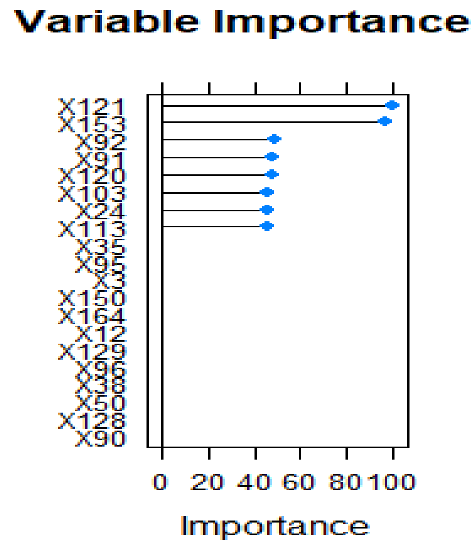

Figure 4.9 Importance Attributes

The chosen attributes from features selection is 20 attributes as shown on Figure 4.8 while in the Figure 4.9 showed the importance percentage of each attributes. As we can see the most important attribute with 100% importance is X121.

3. Fuzzy C-Means

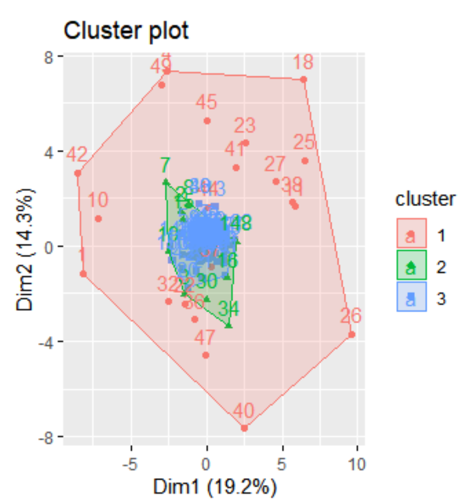

Figure 4.10 Cluster Plot of FCM

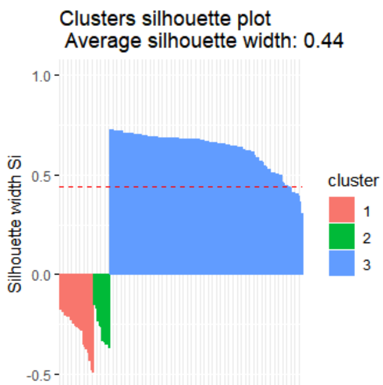

Figure 4.11 Silhouette Plot of FCM

5. Hard C-Means

Figure 4.12 Cluster Plot for HCM

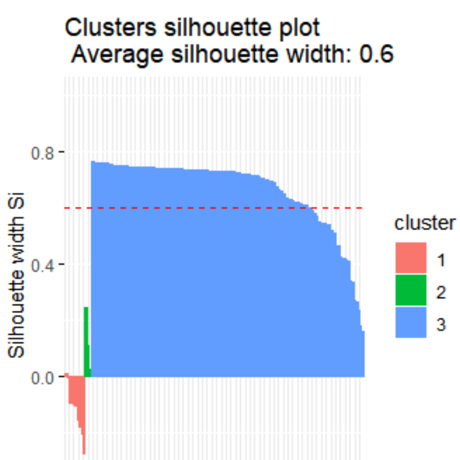

Figure 4.13 Silhouette Plot of HCM

For the overall results of FCM and HCM, we can conclude that this dataset is more suitable for HCM. This is because, the average of silhouette for fuzziness of the dataset is 65%. As we can see, the average of silhouette for HCM (60%) is closer to fuzziness of the dataset compared to FCM (44%).

V. CONCLUSION

The proposed clustering and classification using Fuzzy C-Means (FCM) and Hard C-Means (HCM) for epileptic seizure recognition is an additional value added service module to be adopted by data scientist as their advanced analytics tools to increase accuracy of epileptic seizure recognition. The clustering and classification with FCM and HCM analytics offers a mean of visualization on the accuracy of prediction. With this visualization, the result of patient after been analyze whether they have epileptic seizure or not will be more accurate. However, there are room for improvements for this model in the future, this is because the current project lacks a holistic training data set. It is believed that the whole approach is on the right track, and it can yield better and more accurate result given a proper training data set.

ACKNOWLEDGEMENT

We officially acknowledge Kaggle.com as the data provider for the accurate prediction on Epileptic Seizure Recognition dataset using HCM and FCM algorithm.

REFERENCES

[1] Harun Ur Rahid. (2018). Epileptic Seizure Recognition. Retrieved from Kaggle, https://www.kaggle.com/harunshimanto/epileptic-seizure-recognition

[2] R. J. Almeida and J. M. C. Sousa, "Comparison of fuzzy clustering algorithms for Classification", International Symposium on Evolving Fuzzy Systems, pp. 112-117, 2006.

[3] A.K. Jain, R.C. Bubes, Algorithm for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[4] M. S. Yanp, K.L. Wu and J. Yub, “A novel fuzzy clustering algorithm”, IEEE International Symposium on Computational Intelligence in Robotics and Automation, vol. 2, pp. 647- 652, 2003.

[5] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57

[6] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York

[7] Mayo Clinic Staff. (2019). Epilepsy Symptom. Retrieved from https://www.mayoclinic.org/diseases-conditions/epilepsy/symptoms-causes/syc-20350093

[8] WebMD. (2019). Understanding Seizures and Epilepsy. Retrieved from https://www.webmd.com/epilepsy/guide/understanding-seizures-and-epilepsy#1

[9] Epilepsy Society. (2019). A Closer Look at EEG. Retrieved from https://www.epilepsysociety.org.uk/closer-look-eeg#.XgIXrkczbIU

[10] Steven C. Schachter. Joseph I. Sirven. (2013). Retrieved from https://www.epilepsy.com/learn/diagnosis/eeg/how-read-eeg

[11] Ana Coluby. (2015). Fuzziness in data analysis: Towards accuracy and robustness. Retrieved from https://www.sciencedirect.com/science/article/pii/S0165011415002353

[12] Sejal Jaiswal. (2018). K-Means Clustering in R Tutorial. Retrieved from https://www.datacamp.com/community/tutorials/k-means-clustering-r

[13] Gabriel Martos. (2017) Cluster Analysis with R. Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html

[14] Zeynel Cebeci. (2017). Partitioning Cluster Analysis Using Fuzzy C-Means. Retrieved from https://cran.r-project.org/web/packages/ppclust/vignettes/fcm.html

[15] Binoda, Mohit, Geeta, Tarandeep. (2016). Performance Measure of Hard c-means, Fuzzy c-means and Alternative c-means Algorithms. Retrieved from https://pdfs.semanticscholar.org/dff2/ca51a128b44cd88686f01b5d62f0f1916b68.pdf

[16] Quan Wen, M Emre Celebi. (2011). Hard versus fuzzy c-means clustering for color quantization. Retrieved from https://link.springer.com/article/10.1186/1687-6180-2011-118