[[1]](#footnote-1)

A Study of Hard C-Mean and Fuzzy C-Mean Clustering Method on Diabetes Dataset

Quek Yao Jing, Teoh Kah Lun, Tea See Nai, Ang Kian Chong

*Abstract*—The blooming of machine learning and deep learning is starting to raise the demand for using data to solve our real-world problem. But not all the data collected is crisp, which may contain fuzziness. Unsupervised clustering in a fuzzy logic or fuzzy c-mean is one of the most commonly used methods in data mining when dealing with uncertainty or vague dataset. In this paper, we will study the difference between the clustering method of using the Hard C-Mean and Fuzzy C-Mean method. We will highlight the best cluster found by HCM and FCM.

*Keywords*—Fuzzy Logic; K-Mean Clustering; Fuzzy C-Mean clustering; Unsupervised learning;

# INTRODUCTION

T

he world is changing fast to coop up with the Industrial 4.0 era, the speed of computing power is faster than before and ways of efficient compared to ten years back. Hence, this leads to the blooming of Machine Learning and Deep Learning, which use data to make the predictive analysis .

So, it is essential for all data scientists to understand the nature of the data. A lot of data scientists believed that data exploration is vital before the model training and deployment. But in real life, there are few data known as the fuzzy dataset [1].

Maria, Jan and Bart believed that instead of exact and strict boundaries, representing data more vaguely might be convenient when dealing with a specific dataset [1]. Therefore, it is crucial that we know what kind of data we are handling before further down to model training.

Fuzzy C-Mean is one of the unsupervised learning methods that able to find out how many clusters in the dataset. So, in our paper, we will be interested in finding out the difference between fuzzy c-mean and hard c-mean and find out whether the result is the same.

# RELATED WORK

FCM is a soft algorithm clustering fuzzy data in which an object is not only a member of a cluster but a member of many clusters in varying degrees of memberships well.In this way, objects located on boundaries of clusters are not forced to belong to a certain cluster fully, but instead, they can be a member of many clusters with a partial membership degree between 0 and 1 [2]. Therefore, in the cases that we cannot easily decide what objects belong to only one cluster, especially with the datasets having noises or outliers.

The previous research state that sometimes KM algorithm may be a good option for exclusive clustering but FCM give better results with overlapping clusters.

# overall project findings

## Data Exploration

(Kah Lun and Ang Kian Chong)

(See Nai)

Briefyly describe abou the diabetes data. How many attributes and rows? Is there any missing data?

How many overlapping?

How much fuzzy?

Between and within class?

Fuzzy partition coefficient?

## Hard c-means

Hard C-Mean is also known as hard clustering. A very well-known method is known as K-Mean clustering. K-Mean Clustering is used to classify data in a crisp sense. This method will only classify a data point to only one cluster.

The first step of computing k-mean is we need to determine the number of the cluster we desired. Next, we will need to assign each of the points to the closest centroid, and each of the collection of the points assigned to the centroid is considered as one cluster. After that, the centroid of the original cluster will be kept on updating based on the points assigned in the cluster. The steps above are repeated until the centroid didn’t change or remains constant.

In the K-Mean clustering algorithm, the data points are cluster together based on the similarity. Typically, we will use euclidean distance to find out squared of the distance between each variable.

(1)

## Fuzzy c-means

We had review how hard c-means or k-mean, now we will be review how fuzzy c-means work. Fuzzy clustering also known as soft clustering is based on the Zadeh’s idea which introduced in 1965 [3].

Unlike k-means clustering that cluster the datapoint to crisp set which is 0 and 1, fuzzy c-means algorithm assigned vectors to all the cluster with the membership value at the interval 0 to 1.

FCM extends crisp classification into fuzzy classification with degrees of membership. Every data point will be assigned with membership value. In which a fuzzy membership matrix is given by:

(2)

The equation at (2) indicates that the sum of membership of an individual across all classes is 1. Besides, it also shows that classes are not empty because one individual belongs to each class. At last, the equation also shows that class memberships are let to be partial and can take on any value between 0 and 1.

To find out the best partition set, the objective function (3) is being used.

(3)

K is the number of the clusters, n is the number of data points. The expression of shows the distance between the individual i and the class center j.

# Results and findings

## Hard c-means and k-means

To verify the accuracy of the results, we can use within and between class scatter matrix to further verified the accuracy. The best cluster suppose to have a low within class and high between class.

The relationship between the number of clusters and winthin cluster Sum of Squares (WCSS) is computed in order to further proving the hypothesis on the best clusters above [4]. The WCSS is denoted as (4):

(4)

The equation (4) denoted that as the points in the clusters while is the cluster centers . K-Means algorithm aims to choose cluster centers that able to minimize the WCSS which is also the inertia or within cluster sum-of-squares.

The figure 1.0 shows the elbow method that tried to get WCSS converge from cluster 2 to 30. We can clearly see that, at cluster 10, the graph is already converge. Hence, we can make a conclusion that the cluster 10 might be the best in Hard-C Mean crisp classification.

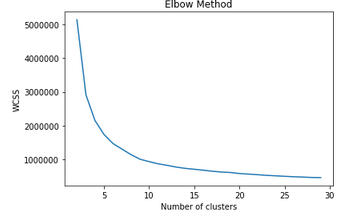


Figure 1.1 shows the Elbow method with WCSS along with cluster 2 to 30.

## Soft clustering

In Soft Clustering or FCM, we can use 2 difference method to find out the best cluster. The first method is fuzzy Partition Coefficient (FPC). FPC has values directly dependent on the relative overlap between non-empty fuzzy clas intersection which denoted as (5):

(5)

Where U is the fuzzy partition matrix being segregated into c classes (partitions), n is the number of dataset, and operation “\*” is a standard matrix multiplication. FPC has special properties:

* if the partitioning is U is crisp (comprising 0 and 1)
* if all the values (complete ambiguity)
* (General)

As the partition coefficient approaches a value of unity, the fuzziness in overlap in classes is minimized. Hence, as increases, the decomposition of the data sets into the classes chosen is more successful.

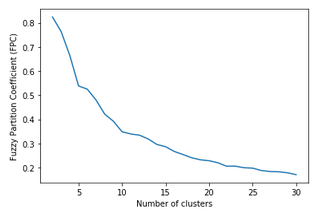


Figure 1.2 shows the Fuzzy Partition Coefficient (FPC) with cluster from 2 to 30.

We had review how fuzzy partition coefficient play its role in choosing the best cluster. So, we will move to another method which is to find out the Intra-cluster compactness, Inter-cluster separation, Inter-cluster overlap measure and Validity Index.

Normarmally, inter-cluster compactness is measured based on the average variation of data points within the cluster. But in our approaches, we use relative dispersion of data points in all the dimensions.

(6)

The formula (6) measured the dispersion of the data based on the coefficient of variant, standard deviation of cluster and coefficient of variation mention in [6]. The formula basically first compute the dispersion of all the data points in each dimensions. As the value become smaller, its tends to close to each other in respective dimensions. Next, we compute the relative dispersion of each data point in dimensions. Then followed by compute the dispersion of number of clusters in each of the dimensions. Finally, at (6) considers the dimensions in which the data points and clusters have maximum dispersion relatively with other dimensions. Therefore, smaller value of Disp(C,U) indicates the higher compactness within the cluster.

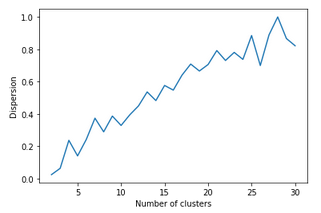


Figure 1.3 shows the dispersion of the datapoints from cluster 2 to 30.

The inter cluster separation is also importance measure for estimate the quality of fuzzy partitions produced by FCM algorithm. The separation is measure by finding the distance between clusters in fuzzy sets. Therefore, we use formula in (7) which is also used in [6] to find our the separation:

(7)

The formula (7) represented as Sep(C,U) considers the clusters with maximum similarity which conversely, results in consideration of the clusters with minimum separation is maximized. Hence, when value of Sep(C,U) increases, indicates the separation of the cluster is better and far from each other.

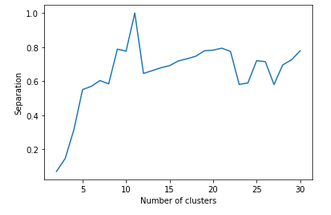


Figure 1.4 shows the separation of the datapoints from the cluster 2 to 30.

On the other hand, we also have Inter-cluster Overlap measure compute overlap of each data points between two fuzzy clusters is represented as in (8) mention by [6]. Degree of overlap is depends on the belongingness of that data point with respect to the clusters and is denoted by . Data that is vague will be assigned a higher value compared to separatable data point. is obtained by summing all overlap data points present in these clusters.

(8)

The fuzzy partitions and value of in which achieve its minimum value will indicate that, the overlap between the highly overlapped pair of clusters is minimum.

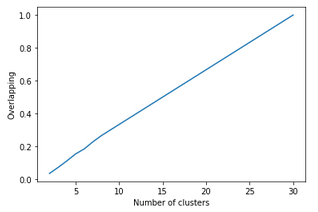


Figure 1.5 shows the overlapping of the datapoints between data points from cluster 2 to 30. From the graph we can see that the overlapping features of the datapoints increases when number of cluster increases.

The last formula used to is the Validity Index which denoted by (9) which is mention by [6]:

(9)

The minimum value of indicates that data points present within clusters are more compact, clusters are overlapped with lesser degree and well separated from each other.

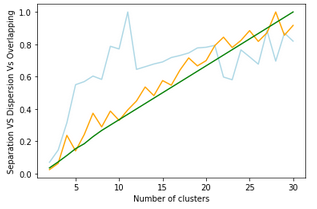


Figure 1.6 shows the graph of separation, dispersion and overlapping.

# Conclusion

See Nai

Why we get 2 difference result? Which one do we need to follow? FCM or normal K-Mean (if Kah Lun find out there are a lot of fuzziness in the data then use FCM, if Kah Lun find out the data no fuzzy use K-Mean).

References

1. [*Maria Óskarsdóttir*](http://www.dataminingapps.com/dma_staff/maria-oskarsdottir/)*,* [*Jan Vanthienen*](https://feb.kuleuven.be/Jan.Vanthienen)*,* [*Bart Baesens*](http://www.dataminingapps.com/dma_staff/bart-baesens/)*.* (May 23. 2016), *[Electronic resources]. There’s Something Fuzzy in Data Mining – Access mode:*

[*http://www.dataminingapps.com/2016/05/theres-something-fuzzy-in-data-mining/*](http://www.dataminingapps.com/2016/05/theres-something-fuzzy-in-data-mining/)

1. *Zeynel Cebeci1, Figen Yildiz, Comparison of K-Means and Fuzzy C-Means AlgorithmsonDifferentCluster Structures*
2. *Salam Al-Augby, Sebastian Majewski, Agnieszka Majewska, Kesra Nermend. A comparison of K-Means and Fuzzy C-Means clustering methods for a sample of gulf cooperation council stock markets.*
3. *Cory Maklin,(Dec 29, 2018), [Electronic resources]. K-Mean Clustering Python Example.*

*Access mode:*

[*https://towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203*](https://towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203)

1. *Ferenc Kovács, Csaba Legány, Attila Babos. Cluster Validity Measurements Technique.*
2. *Neha Bharill, Aruna Tiwari. Enhanced Cluster Validity Index for the Evaluation of Optimal Number of Clusters for Fuzzy C-Means Algorithm.*

1. [↑](#footnote-ref-1)