Fuzzy C-Means

Andrei Paraschiv - Data Mining, Sem 1, 2018-2019 / Master ACS

1. Introduction

Fuzzy C-Means is an unsupervised clustering method developed by Dunn in 1973 [1] and improved by Bezdek in 1981 [2]. Clustering is one of the most important unsupervised learning problems - finding a structure in a collection of unlabeled datapoints.

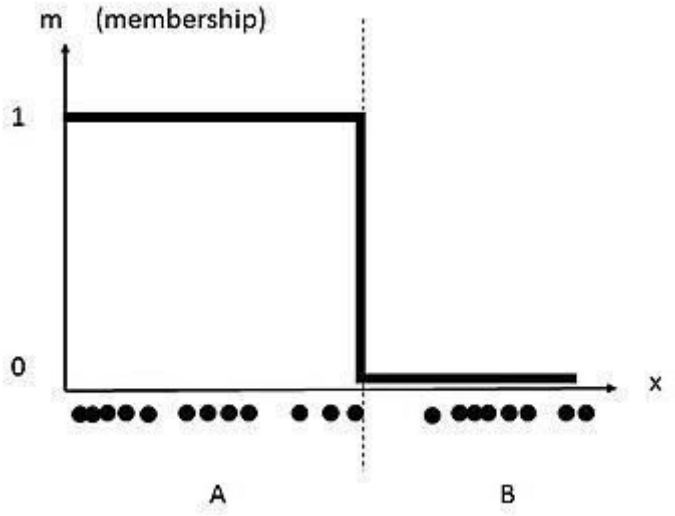
Clustering can be viewed as the process of organizing objects or events into groups of members that share a high number of similarities.

The goal of Fuzzy C-Means clustering is to determine the intrinsic grouping of a dataset by allowing a datapoint / object / event to have a degree of membership into each group. Any individual datapoint can belong to two or more clusters based on a membership function. In real applications very often there are no sharp borders between groups, this is why a fuzzy clustering method is often preferred to a hard clustering one and the results can identify better the internal structure of the data.

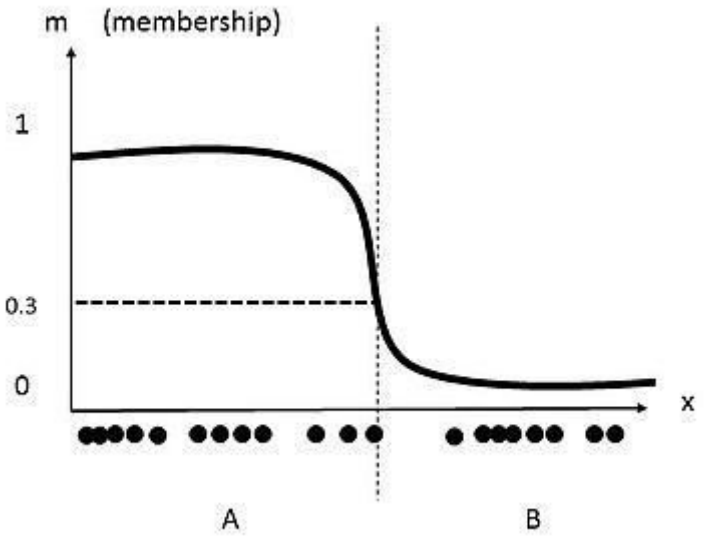
2. Algorithm description

Fuzzy C Means algorithm is a clustering method, an extension of K Means, that allows one data point to belong to two or more clusters based on a membership function. This clustering method is frequently used in pattern recognition.

In a hard clustering algorithm, the membership function the membership function takes the shape of a rectangular function



while in a fuzzy clustering method we might have a membership function similar to a sigma shaped function



The goal of the FCM algorithm is to minimize the objective function

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^m \| x_i - c_j \|^2$$

where w_{ij} is the degree of membership for the datapoint x_i to the centroid c_j and m>=1 is a hyperparameter for the clustering which determines the fuzziness of the clusters.

The centroids are computed using :

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}$$

and the membership degrees :

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\left(\frac{2}{m-1}\right)}}$$

In order to start we need to choose a number of clusters and randomly assign data points to this clusters.

For each iteration we compute the centroids for each cluster, for each data point we compute the membership degree for each cluster based on these new centroids.

We repeat the iterations until the maximum change in membership degrees between two iterations is less then the sensitivity threshold ε

To summarize, the algorithm steps are:

1. Select an initial fuzzy partition by random (assign the values for all w_{ij})

2. compute the centroids for each cluster using the fuzzy partition
3. update the fuzzy partition (the values for w_{ij})
4. repeat step 2 and 3 until the centroids do not change or the change is below a specific threshold ε

2. Importance and practical applications

Clustering is an important classification method. FCM, being an unsupervised machine learning algorithm with an relative low order of complexity, it can be easily used for clustering large datasets with good results.

Due to the nature of fuzzy logic that deals with approximate and non-precise logic, FCM is more suited for coping with the nature of reality then hard clustering methods.

There are many usages for FCM in data analysis, pattern recognition, image segmentation.

Main areas of implementation are

- *Marketing*: customer segmentation, finding customers with similar behaviour or interests
- *Biology*: classification of plants, animals
- *WWW*: document classification, web log analysis
- *Insurance*: grouping customers by risk, finding unusual risk pattern, fraud detection
- *City Planning*: grouping houses by value, type and geographical location
- *Medicine:* patient classification, clustering in medical diagnostic systems

3. Pros and Cons

Some of the main advantages of Fuzzy C-Means are:

- The algorithm always converges
- It is unsupervised
- Leads to a more natural approach to clustering
- It has a relative simple implementation
- The implementation is relativ simple

The weak points and issues encountered are

- The complexity is higher than K-Means - $O(ndc^2i)$ vs $O(ndci)$
- It has a sensitivity to the initial guess
- Since all weights for a given point add up to 1 - it has a sensitivity to outliers and noise
- The effectiveness of the method is depending on the definition of the distance

4. Datasets used

a) Weightlifting Dataset

The Powerlifting datasource is a dataset containing powerlifters performance at specific sport meetups.
The features present in this source are:

MeetId, Name, Sex, Equipment, Age, Division, BodyWeightKG, WeightClassKg, Squat4Kg, BestSquatKg, Bench4Kg, BestBenchKg, Deadlift4Kg, BestDeadliftKg, TotalKg, Place, Wilks

From these features i selected Age and BestBenchKg (Best result at bench weight lifting) for the clustering.

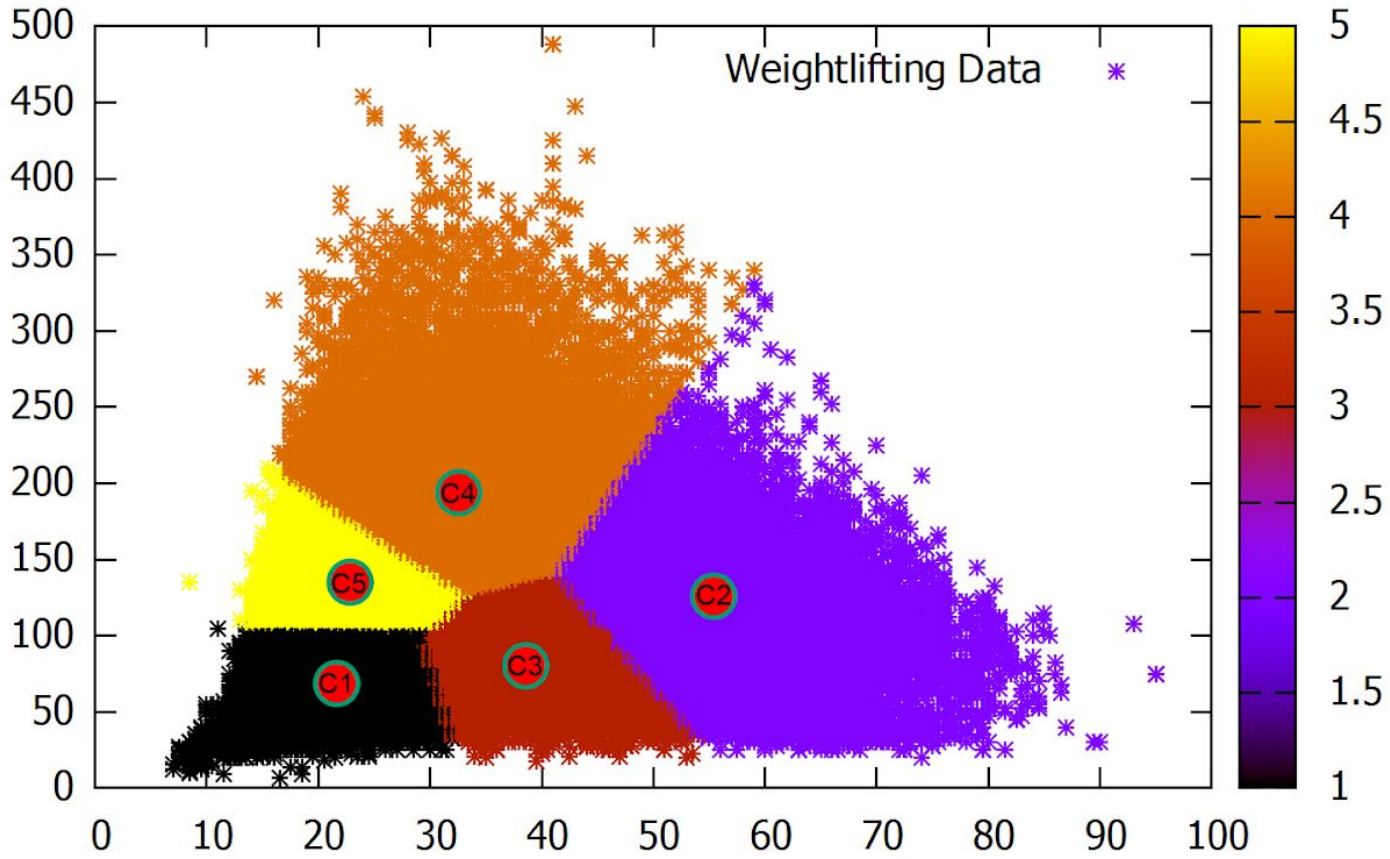
I have run FCM for 5 Clusters, with an epsilon = 0.0001 and maximum 100 Iterations. For the distance evaluation function I have used euclidean distance.

The results for 5 clusters are the following:

Number of Clusters: 5
Number of Samples: 131847

Cluster	Centroids	Number of observations
C1	Age= 21,64 Weight lifted= 68,66	31296
C2	Age= 55,41 Weight lifted= 125,99	17974
C3	Age= 38,58 Weight lifted= 80,30	18908
C4	Age= 32,57 Weight lifted= 193,95	27082
C5	Age= 22,84 Weight lifted= 135,17	36587

Cluster distribution:



As we can see, the high concentration is in clusters C5 and C1, which is confirming our intuitions that many of the competitors are of younger age. We can also see the compactness of those two clusters, these competitors are very close in performances.

The most interesting cluster is C4 where we can see that competitors begin to emerge as clear outliers / winners

b) Income Tax data NY

The income Tax dataset contains all income tax returns by companies in New York between 2001 and 2012.

The features present in this source are:

Tax Year, Resident Type, Place of Residence, Country, State, County, Income Class, Disclosure, Number of All Returns, NY AGI of All Returns, Deductions of All Returns, Dependent Exemptions of All Returns, Taxable Income of All Returns, Tax Before

Credits of All Returns, Tax Liability of All Returns, Place of Residence Sort Order,Income Class Sort Order

From these features i selected Aggregated Gross Income of all Returns (AGI) and Taxable income of all returns (TI) for the clustering.

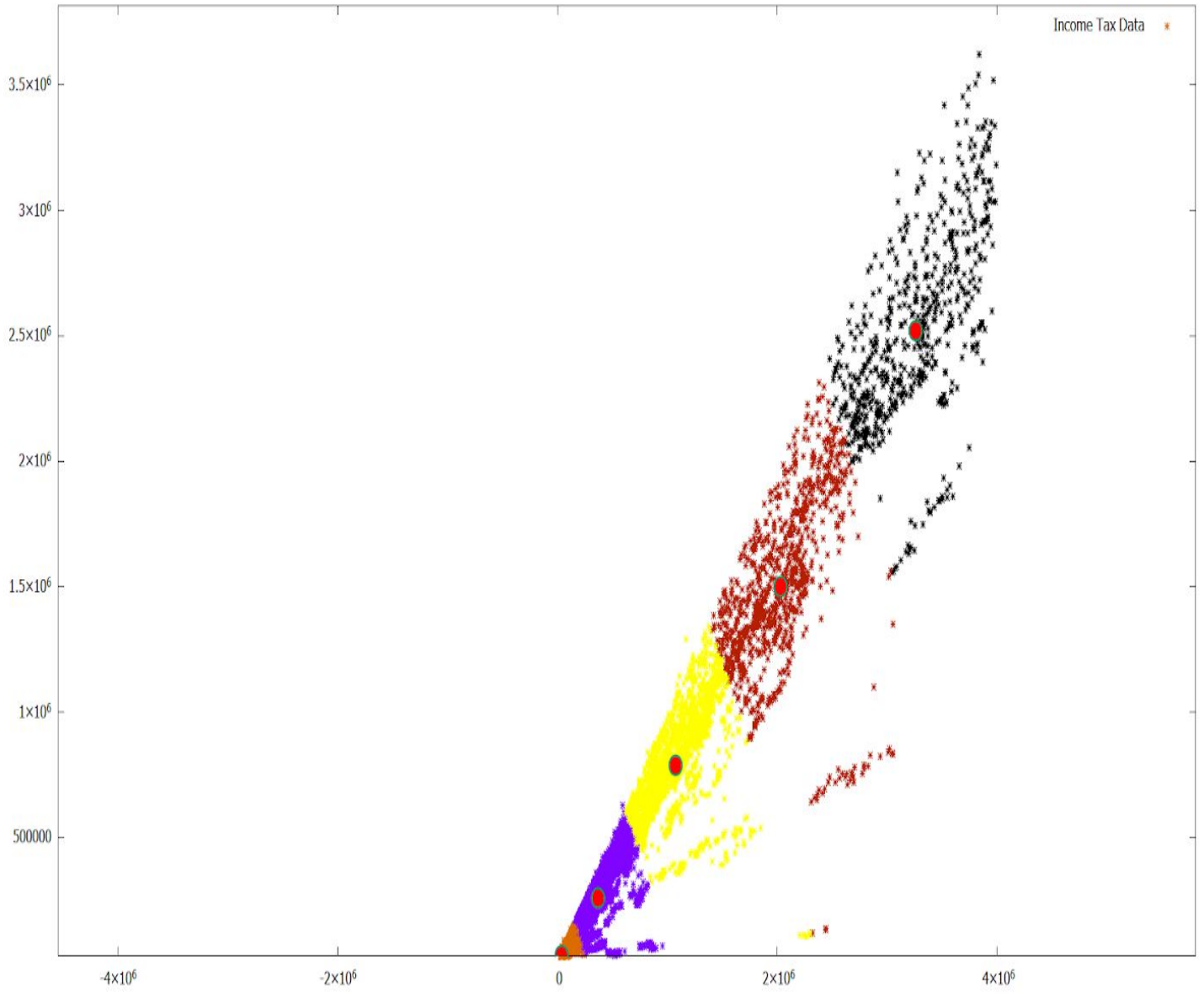
I have run FCM for 5 Clusters, with an epsilon = 0.0001 and maximum 100 Iterations. For the distance evaluation function I have used euclidean distance.

The results for 5 clusters are the following:

Number of Clusters: 5
Number of Samples: 21693

Cluster	Centroids	Number of observations
C1	AGI = 21,64 TI= 68,66	606
C2	AGI = 21,64 TI= 68,66	3443
C3	AGI = 21,64 TI= 68,66	904
C4	AGI = 21,64 TI= 68,66	15130
C5	AGI = 21,64 TI= 68,66	1610

Cluster distribution



We can see a concentration in Cluster C4 and a very scattered cluster C5. The intuitive explanation for cluster C4 is the presence of middle size companies in that cluster. As we know, small and middle size companies are the most in a national economy

5. Results

In my implementation, Fuzzy c-means clustering did perform well in regards of speed - having runtime of 10 s for 130.000 records on a laptop with intel i5 and 4 gb ram
The clusters matched our intuition, grouping datapoints in well defined categories
As in many data mining applications, the quality of the results depend on the initial cleansing of data and normalization. Even though fuzzy c-means is sensitive to outliers, due to the amount of data in the datasources those were suppressed, and i did not encounter distortions
Some practical usages for the found partitions could be:

- Detecting doping cases by selecting outliers and datapoints with high distance from the centroids

- Creating a fair evaluation scale for weightlifters by ranking them inside their cluster versus global ranking
- Detecting high risk clusters of tax evaders or fraudulent activities
- Detecting money launderers by detecting outliers and datapoints far from the cetroids

6. References

[1] J. C. Dunn (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics, 3:3, 32-57, DOI: 10.1080/01969727308546046
[2] Bezdek, James. (1981). Pattern Recognition With Fuzzy Objective Function Algorithms. 10.1007/978-1-4757-0450-1.
[3] Wikipedia, https://en.wikipedia.org/wiki/Fuzzy_clustering
[4] Guijarro Rodríguez, Alfonso & Cevallos, Lorenzo & Botto-Tobar, Miguel & Leyva-Vázquez, Maikel & Yepez Holguin, Jessica. (2017). Clinical Assessment Using an Algorithm Based on Fuzzy C-Means Clustering. 181-193. 10.1007/978-3-319-67283-0_14.