CSC730: Report for Assignment 2

South Dakota School of Mines and Technology

Jacob James, John Nelson, Kris Jensen

January 28, 2024

1 Introduction

The second assignment of this course tasked us with implementing a Fuzzy C-Means (FCM) clustering algorithm 'from scratch' on a set of skewed data from the MNIST dataset. This skewed dataset contained eight classes from a total set of ten classes. Also, each class was not represented with equivalent frequency. The dataset contains 12244 records of data. Each record contains a 784-element list that represents a 28x28 image of a handwriting sample.

This report will detail the steps taken by our team to implement the FCM algorithm 'from scratch' and determine the accuracy using the Rand Index Metric.

All code was executed in python on VS code.

2 Methodology

Implement fuzzy-cmeans from scratch

The paper this will be implemented from is titled, *A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis*

From reference [1], the FCM algorithm is defined as follows:

> FCM pseudo-code:
> Input: Given the dataset, set the desire number of clusters c, the fuzzy parameter m (a constant > 1), and the stopping condition, initialize the fuzzy partition matrix, and set stop = false.
> Step 1. Do:
> Step 2. Calculate the cluster centroids and the objective value J.
> Step 3. Compute the membership values stored in the matrix.
> Step 4. If the value of J between consecutive iterations is less than the stopping condition, then stop = true.
> Step 5. While (!stop)
> Output: A list of c cluster centres and a partition matrix are produced.

The equations that are used in the FCM algorithm are defined as follows:

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^m \underline{p_i}}{\sum_{i=1}^{n} \mu_{ik}^m} \tag{1}$$

$$|\underline{p_i} - \underline{v_k}| = \sqrt{\sum_{i=1}^{n} (\underline{x_i} - \underline{v_k})^2} \tag{2}$$

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^m |\underline{p_i} - \underline{v_k}|^2 \tag{3}$$

$$\mu_{ik}^m = \frac{1}{\sum_{l=1}^{c} \left(\frac{|\underline{p_i} - \underline{v_k}|^2}{|\underline{p_i} - \underline{v_l}|^2}\right)^{\frac{2}{m-1}}} \tag{4}$$

The first step in the psuedocode sets up a loop to run until the stop condition is met. The stop condition is met when the value of J between consecutive iterations is less than the stopping condition. The second step in the psuedo code requires generating a random μ matrix, then calculating the cluster centroids from eq. 1 and the objective value J from eq. 3. The third step in the psuedo code requires calculating the membership values from eq. 4. The fourth step in the psuedo code requires checking the value of J between consecutive iterations. If the value of J between consecutive iterations is less than the stopping condition, then stop = true. The fifth step in the psuedo code requires looping back to step 2. The output of the psuedo code is a list of c cluster centres and a partition matrix.

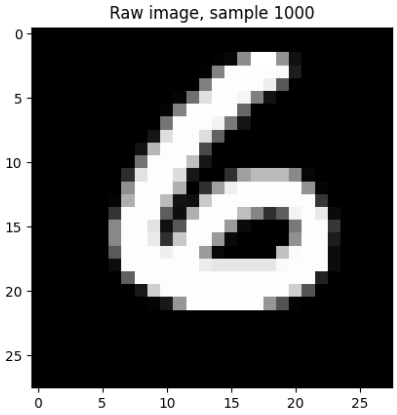

Figure 1: Handwriting Sample 1000

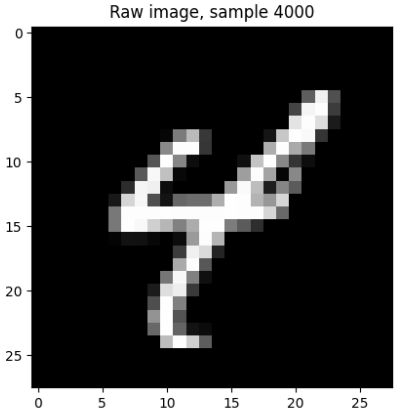

Figure 2: Handwriting Sample 4000

An example of two handwriting samples is shown as images in Figure 1 and Figure 2. These images were produced using the 'imshow' function from the 'matplotlib' Python library.

Our group independently implemented the FCM algorithm in python then compared results among the group members. After comparison of results, we determined corrected issues where necessary. The final implmentation of the FCM code has been attached via D2L in an appropriately named Jupyter notebook.

The FCM algorithm was ran against two versions of the skewed MNIST dataset. The first version retained the 784 dimensions, while a second version utilized principal component analysis to reduce the dimensionality of the dataset to 2 dimensions. The results of the FCM algorithm were compared to the actual labels of the dataset using the rand index. The rand index is a value between 0 and 1. A value of 1 indicates that the clusters are identical to the labels. A value of 0 indicates that the clusters are completely different from the labels. The rand index is a good measure of the accuracy of the clustering algorithm.

3 Results

One requirment of our assignment was to review the clustering accuracy of the fuzzy c-means algorithm as applied to the skewed MNIST dataset. The fuzzy c-means algorithm is a clustering algorithm that is based on the minimization of the objective function. The objective function is a function that is used to measure the quality of a clustering. The objective function was defined above in equation 3. When clustering is finished we need to compare the clusters that were produced to the actual labels of the data. We will use the rand index to compare the clusters to the labels. The rand index is defined as follows:

$$RI = \frac{n_{00} + n_{11}}{n_{00} + n_{11} + n_{01} + n_{10}} = \frac{n_{00} + n_{11}}{\binom{n}{2}} \tag{5}$$

where n_{11} is the number of pairs of elements that are in the same cluster and in the same class, n_{00} is the number of pairs of elements that are in different clusters and in different classes, and $\binom{n}{2}$ is the total number of pairs of elements in the dataset. The rand index is a value between 0 and 1. A value of 1 indicates that the clusters are identical to the labels. A value of 0 indicates that the clusters are completely different from the labels. The rand index is a good measure of the accuracy of the clustering algorithm.

While the FCM algorithm was running, we decided to review the progression of J, eq. 3.

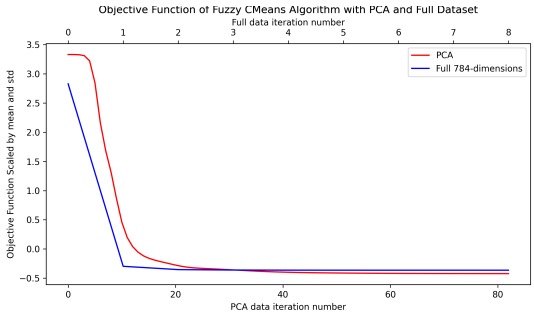

Figure 3: St. Dev. Normalized obejective functions through runtime of the FCM algorithm

The RI results for the full data set resulted in a value of 0.692 and the PCA dataset resulted in a value of 0.672.

We also plotted the results of FCM applied to the PCA dataset. The clusters are graphed according by color according to their class label. There is also a black X to indicate the cluster center.

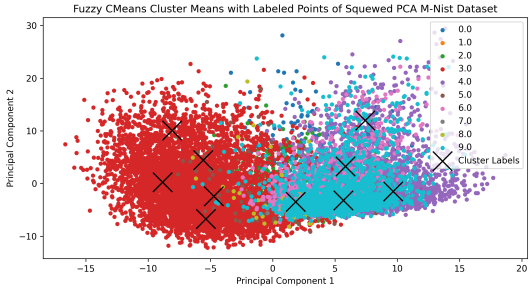

Figure 4: Clusters produced by FCM applied to the PCA dataset

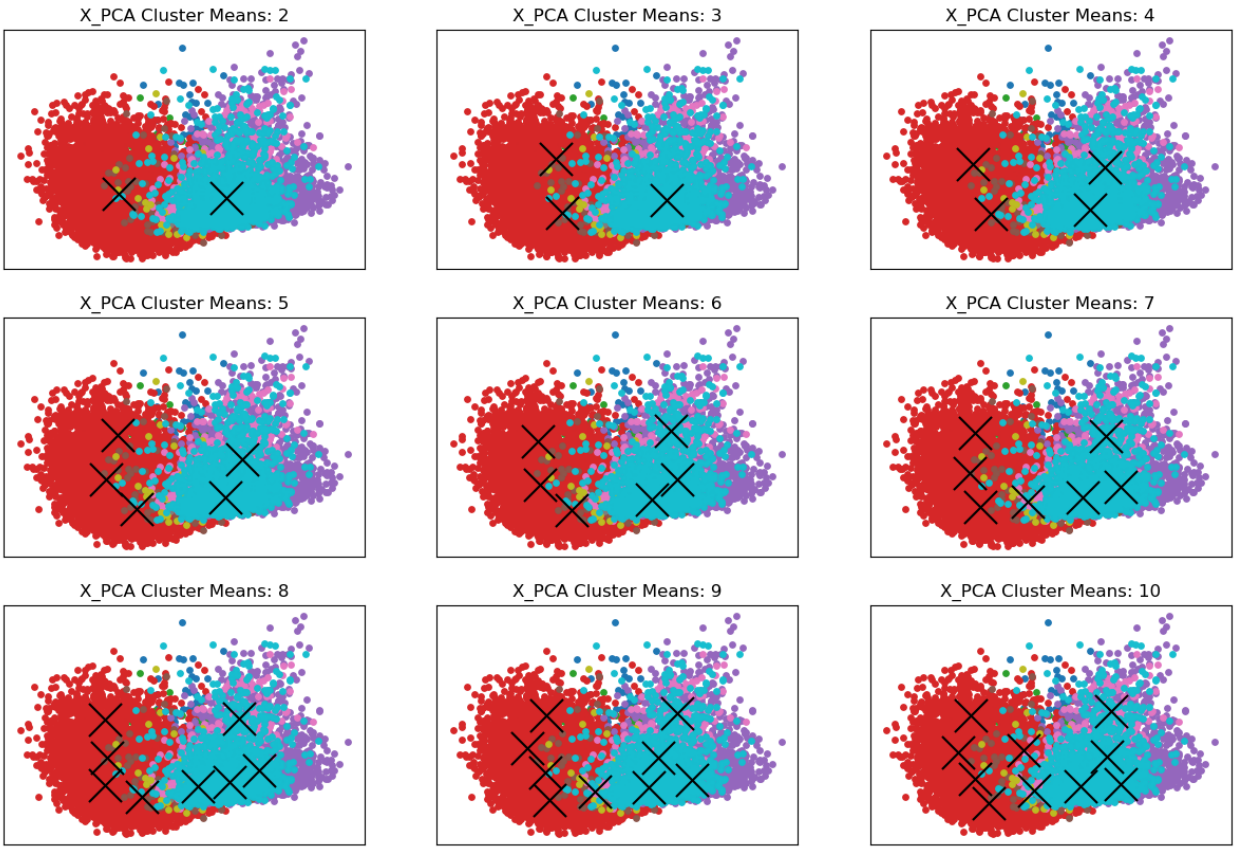

Figure 5: RI results for different cluster amounts applied to the PCA dataset

Additionaly, we wanted to evaluate the PCA dataset with different cluster amounts. We ran the FCM algorithm with cluster amounts of 2, 3, 4, 5, 6, 7, 8, 9, and 10. The results are shown in figure 5.

4 Discussion

Typical real world clustering problems run can have subjective clustering numbers. For this reason, an experiment was done to determine the rand index for all integers in $[2, 10]$ cluster means in order to determine the optimal cluster amount which represents our particular sqewed MNIST dataset. Using the histogram in figure 6, we can estimate the appropriate amount of clusters which would perform the highest Rand Index, and determine the validity of our hypothesis with our experiment. Given the results discussed in the Results section, it appears that four clusters in the PCA representation and 3 clusters in the fully dimensioned representation of the sqewed MNIST gave the best rand index score, even though ten unique labels were used in the truth labels. In the future, this experiment can be run with different sqewed representations of the MNIST dataset to determine if the results are consistent.

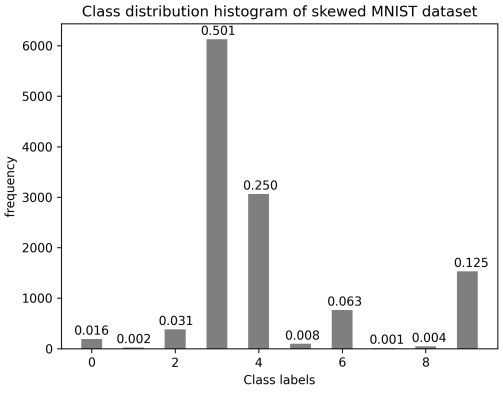

Figure 6: Class Representation in Sqewed MNIST Dataset

Additionally, consideration must be taken regarding the rand index metric on a soft clustering algorithm like FCM. The implementation of the rand index in this assignment took the maximum probability a point would be within a given k^{th} cluster, and then implemented the rand index as described in the methods section. The rand index calculates the similarities of points in different clusters as a fixed yes or no. As FCM is a soft clustering algorithm, the probabilities of points being within all clusters are returned, therefore a different metrics could potentially give more insight to the accuracy of soft clustering algorithms.

Lastly, two representations of the data was evaluated: PCA and Scaled. PCA scales down the dimensionality of the data to two dimensions which theoretically decreases the complexity, but also hinders the accuracy of the clustering method. In this implantation, the reduction of accuracy was only one percent, which inclines us to believe that the PCA representation would be sufficiently similar to a fully dimensional representation on a larger dataset with similar sqewed ratios.

5 Conclusion

The Fuzzy C-Means algorithm was successfully implemented "from scratch" on the sqewed MNIST dataset and the rand index metric was used to determine the accuracy. Additionally, the algorithm was run using initial cluster amounts $[2, 10]$ inclusive and determined that the rand index was relatively similar for all cluster amounts within the domain, with slight increase in three clusters in PCA, and four clusters for the all dimension representation of the sqewed MNIST dataset. Future work can be done to determine how different sqewed representations of the MNIST dataset will produce similar or different optimum cluster amounts.

6 References

[1] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267–279.