CSC730: Report for Assignment 2

South Dakota School of Mines and Technology

Jacob James, John Nelson, Kris Jensen

January 28, 2024

1 Introduction

The first 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.

An example of two handwriting samples is shown as images in figure **??** and figure **??**. 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.

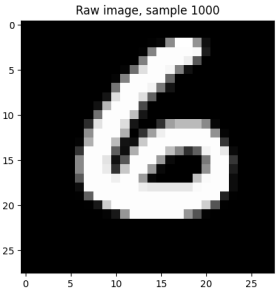

Figure 1: Handwriting Sample 1000

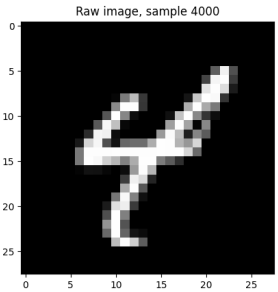

Figure 2: Handwriting Sample 4000

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:

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

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 C_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. The rand index is defined in the scikit-learn library as the adjusted rand index. The adjusted rand index is defined as follows:

$$ARI = \frac{RI - Expected(RI)}{max(RI) - Expected(RI)} \tag{6}$$

4 Discussion

¡insert text here¿

5 Conclusion

¡insert text here¿

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.