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 analyzing 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 generate an anomaly score for each image and compare our results for accuracy.

We used two primary toolsets to perform the analysis. One toolset was python executed on VS code and the other was python executed on Google Colab.

2 Methodology

¡insert text here¿

Implement fuzzy-cmeans from scratch

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

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

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

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

Step 3. Compute the membership values stored in the matrix.

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

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.

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.

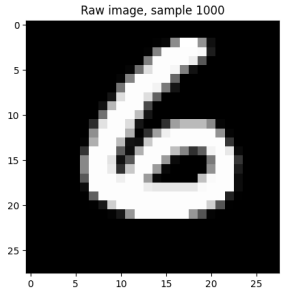

Figure 1: Handwriting Sample 1000

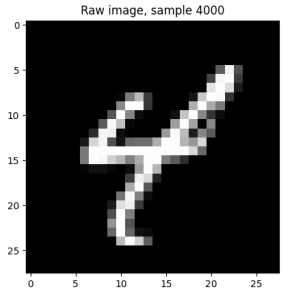

Figure 2: Handwriting Sample 4000

3 Results

¡insert text here¿

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.