Report

Introduction

In this study, we explore the efficacy of utilizing clustering techniques alongside logistic regression to improve classification performance in semi supervised learning. Our primary focus is on the Fashion MNIST and The Overhead MNIST dataset. The main objective is to examine how using representative images from clusters as training data, as opposed to random samples, can enhance the predictive accuracy of a semi supervised learning model.

Data Description

The Fashion MNIST dataset includes 60,000 training images and 10,000 test images spread across 10 classes, including T-shirts/tops, Trousers, Pullovers, Dresses, Coats, Sandals, Shirts, Sneakers, Bags, and Ankle boots. Each image is 28x28 pixels, grayscale. The Overhead-MNIST dataset is a collection of satellite images similar in style to the ubiquitous MNIST hand-written digits found in the machine learning literature.

Methodology

1. **Baseline Model Performance**: A logistic regression model is trained on the entire training dataset and evaluated on the test set.
2. **Clustering and Representative Selection**: The training data is clustered into k clusters using K-Means, and the most central image (closest to the centroid) in each cluster is selected as the representative image.
3. **Training with Representatives**: A logistic regression model is then trained on these representative images and then evaluated on the test set.
4. **Label Propagation**: Labels from the representative images are propagated to all images in the respective clusters, and a model is trained on these labelled data and then evaluated on the test set.
5. **Partial Propagation**: Labels are only propagated to a portion of the data closest to the cluster centroids, which presumably are more representative of the cluster's characteristics and then evaluated on the test set.

The above steps are repeated for various cluster sizes (K=10, 25, 50) to assess the impact of the number of clusters on model performance.

Results

Part 1

Baseline Model Performance

- The logistic regression model trained on the entire dataset served as a benchmark for subsequent techniques.

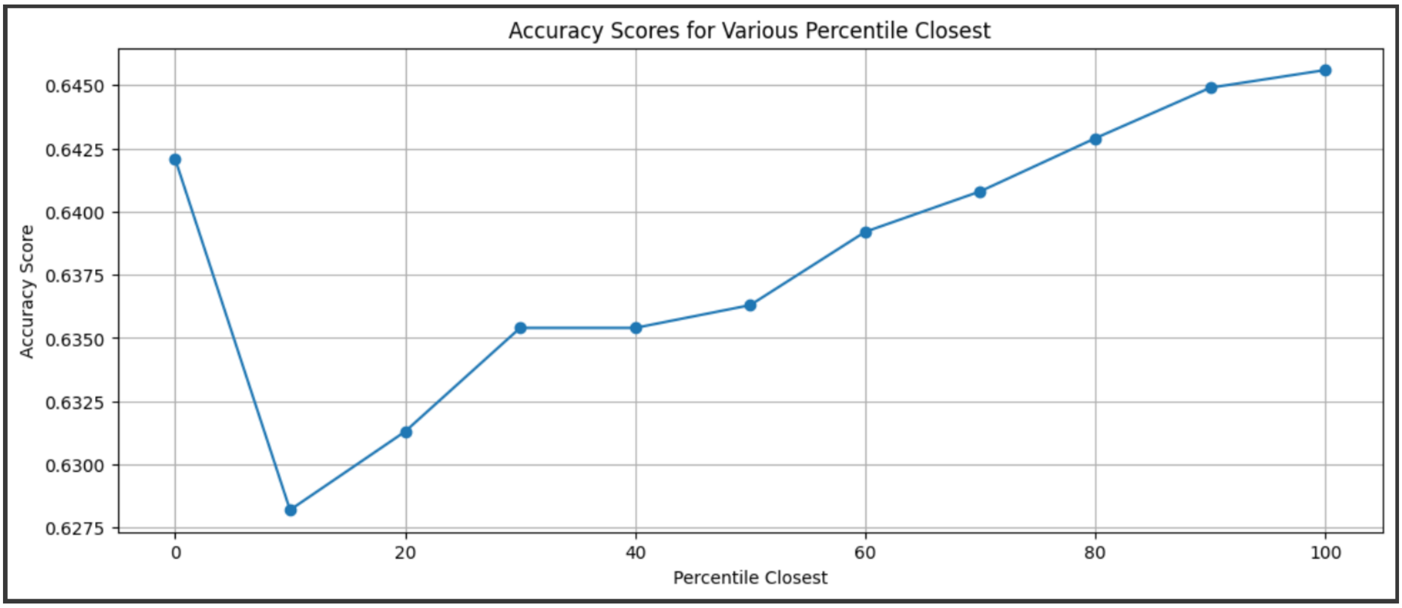
Effect of Clustering

- Training logistic regression models on only representative images (50 instances for each k) yielded lower accuracy compared to the baseline but was better than training on 50 random images.

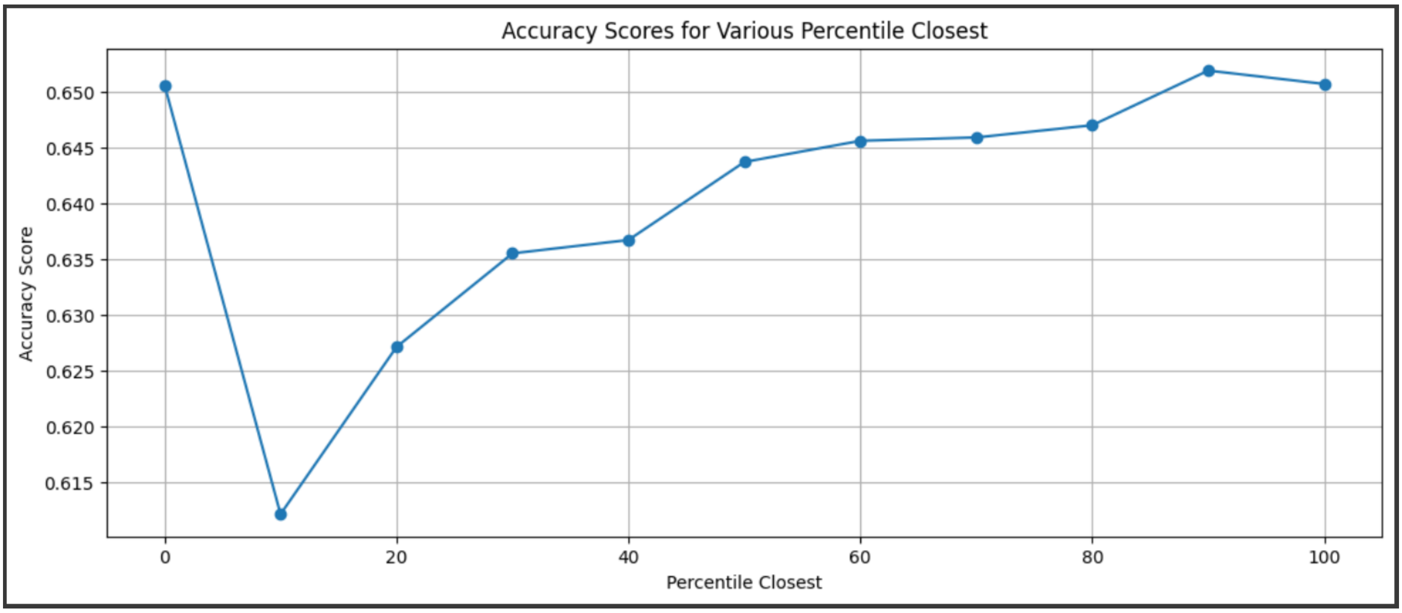
Label Propagation

- Full propagation of labels provided the best model accuracy, demonstrating that more training data, even if somewhat noisy, benefits the model.
- Partial label propagation (propagating labels around to the closest 90% of points in each cluster) generally provided the best results amongst others, particularly at lower cluster counts.

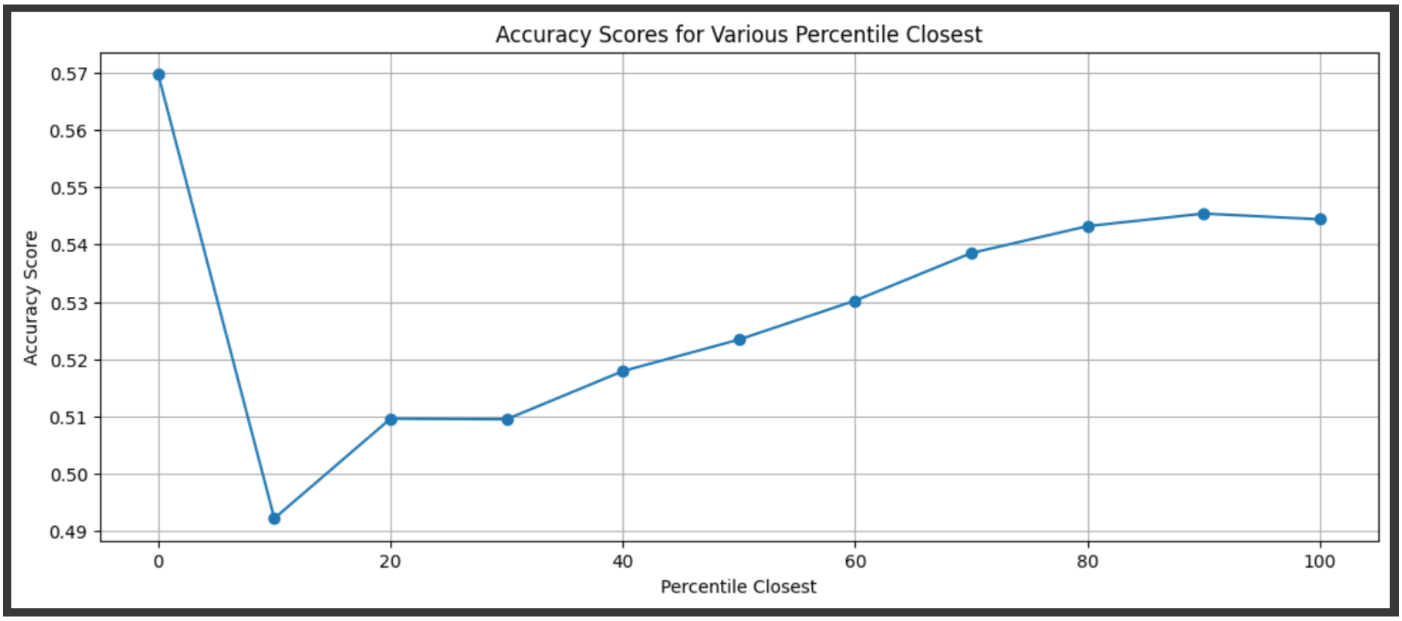
K=50


Accuracy Scores for Various Percentile Closest
Accuracy Score
0.6450
0.6425
0.6400
0.6375
0.6350
0.6325
0.6300
0.6275
0
20
40
60
80
100
Percentile Closest

K=25


Accuracy Scores for Various Percentile Closest
Accuracy Score
0.650
0.645
0.640
0.635
0.630
0.625
0.620
0.615
0
20
40
60
80
100
Percentile Closest

K=10


Accuracy Scores for Various Percentile Closest
Accuracy Score
0.57
0.56
0.55
0.54
0.53
0.52
0.51
0.50
0.49
0
20
40
60
80
100
Percentile Closest

Part 2

Baseline Model Performance

- The logistic regression model trained on the entire dataset served as a benchmark for subsequent techniques.

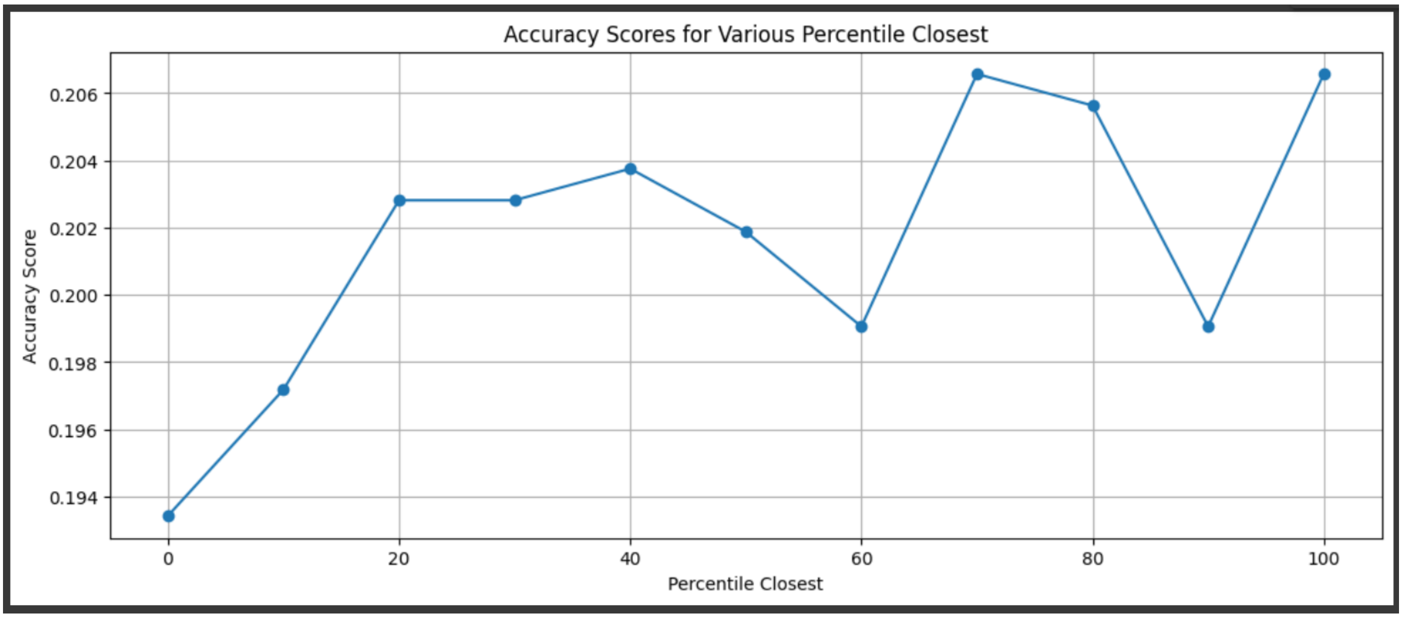
Effect of Clustering

- Training logistic regression models on only representative images (50 instances for each k) yielded lower accuracy compared to the baseline but was better than training on 50 random images.

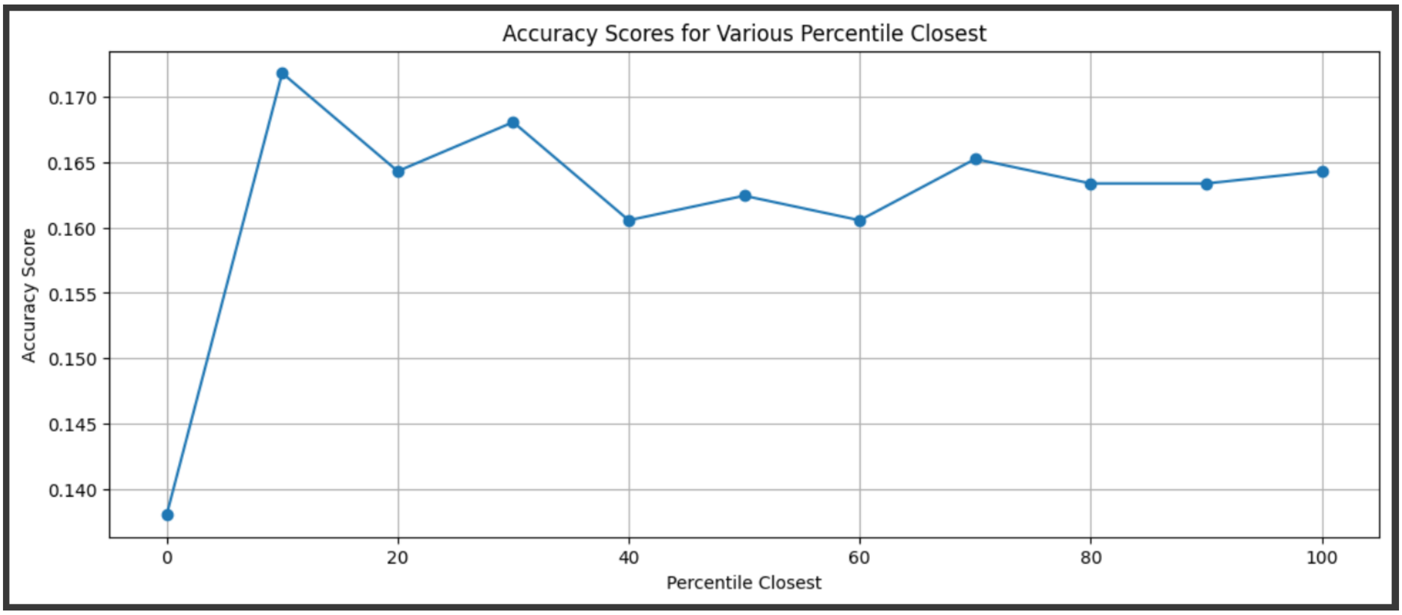
Label Propagation

- Full propagation of labels provided the best model accuracy for higher cluster numbers and worse model accuracy for lower number of clusters.
- Partial label propagation generally provided the best results amongst others, particularly at lower cluster counts. The best results were there when propagating labels only to the closest 70% , 10%, 20% of points for K=50,25,10 respectively.

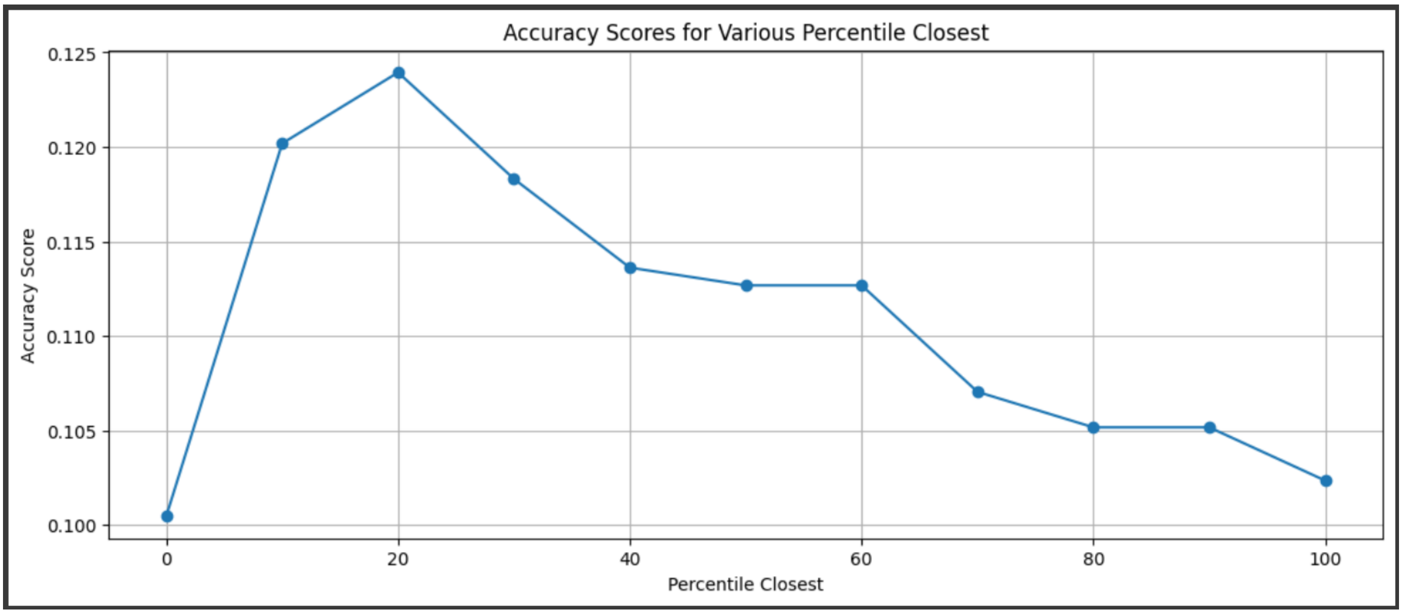
K=50



K=25



K=10



Conclusion

The experiments illustrate that utilizing clustering to select representative training samples can significantly improve logistic regression performance, especially when the training data is limited. Label propagation, particularly selective propagation, further enhances model accuracy, leveraging the structure within the data more effectively than traditional random sampling or even full dataset training in some scenarios.

Submitted by:
Gauranga Kumar Baishya, MDS202325
Dipanjoy Saha, MDS202321