Data Mining and Machine Learning Assignment 3

Moumi Roy – MDS202125

Srijit Saha – MDS202150

Dataset:

We have used the Fashion MNIST which consists of training set of 60000 examples and a test set of 10000 images. Each image is a 28*28 grayscale image. Each pixel represents intensity of the pixel ranging from a value of 0 to 255.

All the examples used in the training set and the test set are assigned to each class label among the following;

T-shirt, Top, Trouser, Pullover, Dress , coat, sandal , shirt sneaker, Bag and Ankle boot.

Objective:

We use K-means clustering to identify a small subset of labeled images to seed the classification process to increase the accuracy.

The main objective is to label the dataset with very few labeled data along with a huge amount of unlabeled ones. For this, we use K-means clustering and label it further by finding an optimal K.
This is an example of semi – supervised learning.

Implementation:

In our example, we have used two training models, mainly logistic regression and K-means clustering, and compared the accuracies between the two, for each case.

Models on the entire training dataset, and marks on randomly labeled K instances of the training set, and also models on the centroids computed using K-means clustering , and propagates labels of the centroid to each data point in the clusters, and then propagates labels to the data points close to cluster centroids.

Libraries used are Numpy , Keras, Tensorflow , sklearn, Matplotlib, time, memory_profiler.

After flattening each grayscale image, we have tested our models subsequently on the flattened data.

We have used various cluster numbers and compared the accuracies of our models across these different cluster numbers.

Listed below are the various clusters that we have taken and the corresponding accuracy for the logistic regression models before and after clustering of the unlabeled data instances.

No. of clusters	Accuracy	Accuracy after clustering
10	0.3736	0.483
30	0.5716	0.635
50	0.6618	0.6631
70	0.691	0.659
90	0.6825	0.6928

Time taken for running the models :- 6303.40 seconds (~1 hr 45 mins),the peak memory is 1057.15 MiB and the increment is 0.05 MiB.

Since the maximal increase in accuracy is seen for cluster k=10, we try to propagate the labels to all the other instances in the same cluster. Accuracy becomes 47.6 %.
But if we do label propagation on closest 25% instances we get an accuracy of 47.37%.

Peak memory: - 1845.43 MiB , Increment = -0.01MiB

We observe that the overall accuracy is not that good which is expected as we use cluster size 10 that means on average 1 example per cluster.

We then experiment for k=100, that is 10 examples per cluster.
Here for logistic we get a accuracy of 69.14%.Upon clustering we get an increment upto 69.86%.After label propagation to all the instances the accuracy becomes 69.44%.
But if we do label propagation on closest 25% instances we get an accuracy of 70.52%.

Conclusion:

We observe that the larger the dataset that could be used, the better the results that we can get in terms of accuracy.

We also see that as a larger number of clusters that we use in labeling the unlabeled instances, centroids computed are more accurate to the true representations of the classes.

We mainly notice that in case we use lower number of clusters in the semi – supervised learning, applying clustering on the data, followed by Logistic regression on the data is better in term of accuracy than simply applying logistic regression.