Method 1:

We first divided the training data into clusters, then labelled the representative images from these clusters and then applied Logistic regression on these labelled images.

The results for this method are as follows:

The accuracy appears to be the best for k=500

Model	Accuracy
Logistic Regression on entire dataset	85.57%
Logistic Regression for 500 clusters' representative images	78.72%
Logistic Regression for labels propagated to entire dataset	77.63%
Logistic Regression for labels propagated to the nearest 10% elements in each cluster	77.30%

Method 2:

Here, we use Principal Component Analysis (PCA) to reduce the number of attributes from 784 to say, 25. Then as before, we do cluster and then Logistic Regression and then we also try Support Vector Machines (SVM).

The results for this method are as follows:

The accuracy seems to be the best for 400 clusters, for logistic regression.

Model	Accuracy
Logistic Regression for 400 clusters on the representative images	75.65%
Logistic Regression for labels propagated to the entire data set	76.19%
Logistic Regression for labels propagated to nearest 10% elements for each cluster	75.91%

300 is the best number of clusters empirically here for SVM

Model	Accuracy
SVM for 300 clusters on the representative images	75.64%
SVM for labels propagated to the entire data set	76.12%
SVM for labels propagated to nearest 10% elements for each cluster	77.48%