Digit recognizer

Digit recognizer design is a classic topic in data mining and machine learning area. The dataset we use comes from the MNIST(Modified National Institute of Standards and Technology). The digit recognizer design is also among the most popular classification topics in Kaggle. Figure 1 shows an example of the handwriting image. The image can be represented with pixels matrix, with each pixel has single integer value between 0-255, inclusively. In our training dataset, the matrix has a dimension of 28 pixels in height and 28 pixels in width. The matrix can be represented with a 1-dimension array with 784 pixel values.

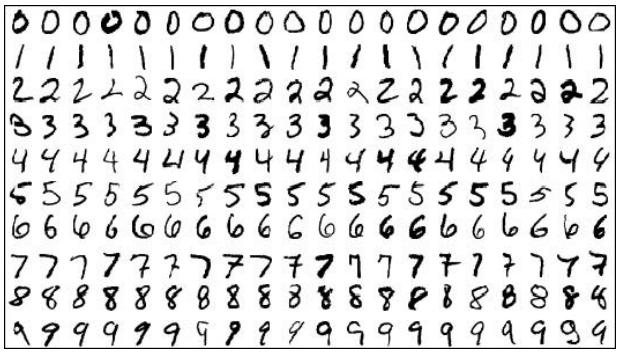

Figure 1: An example of hand written digits

In the training set, except the 784 columns of pixel value, there is one additional “label” value, stands for the true digit number drawn by the writer. There are 28000 images in the training dataset. The test data set is the same as the training set, except that it does not contain the "label" column.

Our target is to build recognizers that are able to recognize the handwriting images from its 784 pixel values with optimized accuracy and efficiency. We propose to implement three algorithm, which include random forest, k-NN and artificial neural networks. We will compare the performance of those recognizer and use summary statistics and visualization techniques to present our findings. We will use Weka/R/python as different tools for training/testing the models.

Group Panda
Han Li
Mingrui Wang
Wei Liu