**[Data description]**

The data set used in this example is from the MNIST database (Mixed National Institute of Standards and Technology database). MNIST database is a large database of handwritten digits that is commonly used for training various image processing systems. The original MNIST data set contains 60,000 digits ranging from 0 to 9 for training the digit recognition system, and another 10,000 digits as test data. To avoid storage and memory problems, in this study, a subset of 1000 training data and 200 test data are randomly selected from the original dataset.

In the training data set, each instance has two components. The first component is target label (0-9). The second one is a 28 pixel by 28 pixel grayscale image of the digit. Each pixel is represented by a floating point number indicating the grayscale intensity at the location. The 28 by 28 grid of pixels is unrolled into a 784-dimensional vector. That is equal to say that the training dataset has 784 attributes. Hence, the training set gives a 30,000 by 784 matrix, where every row is a training example for a hand written digit image.

**[Script Description]**

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Main parameters | Accuracy | Running Time  (second) |
| Decision Tree (DT) | minimum sample split is 12  maximum depth is 10 | 0.66 | 3.3 |
| SVM | kernel is 2-order polynomial | 0.94 | 10.2 |
| Boosted Decision Tree (BDT) | maximum depth is 1  number of estimator is 500 | 0.72 | 37.5 |
| ANN | network size is 784-25-10 | 0.91 | 100.6 |
| K-NN | k = 4 | 0.84 | 3.82 |
| Google Prediction (GP) | - | **0.84** | **94.6** |

|  |  |
| --- | --- |
|  |  |
|  |  |

Performance (accuracy and running time) of Google Prediction API are compared with that of classifiers build up with other supervised learning algorithms as shown in table and figures above. These compared algorithms include: Decision Tree (DT), SVM (Supported Vector Machine), ANN (Artificial Neural Network), Boosted Decision Tree (BDT) and KNN (K Nearest Neighbor). For each classifier, a 10-fold cross validation is implemented for model parameter selection to optimize performance. Running time measured includes both training and testing procedure of each classifier.

The comparisons indicate that accuracy of Google Prediction API is placed in the middle of other five algorithms. SVM and ANN, which have been tested to be suitable for hand written digits prediction, outperform. As for running time, google prediction API is more time consuming than most other algorithms except for Artificial Neural Network. In general, google prediction API provides a fairly good prediction.