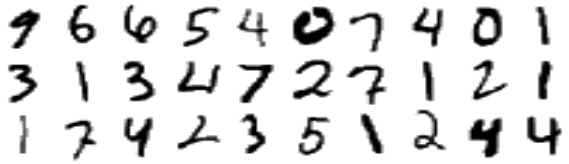
Handwritten Digit Recognition



Project Milestone

Team Members
Akshara Boppidi
Anirudh Nagulapalli
Pooja Jadhav Eshwarlal

I. Abstract

The goal of the project is to take an image of a handwritten single digit and determine what the digit is. This project is chosen from kaggle competitions. The data for this competition is taken from the MNIST dataset (Modified National Institute of Standards and Technology) which consists of training set of 60,000 examples and a test set of 10,000 examples. This data is provided by Kaggle within these data files (.csv) namely train.csv, test.csv.

These data files contain grey-scale images of hand-drawn digits starting from zero to nine. There are 10 digits to predict. Each image contains 784 pixels in total. It is segmented into 28*28 pixels height by width representing the pixels brightness or darkness with a number rating. The rating given to the pixel is between 0 and 255, inclusive. The train.csv contains 785 columns. The first one is the 'label' and is the digit drawn by the user. The rest of the column contains the pixel value of the surrounding pixels.

II. Introduction

Digit recognition is a good problem to learn about machine learning. Some applications for digit recognition are online handwriting recognition on devices and numeric entries in forms filled by hand, investigation, recognition of zip code by postal services. This poses interesting challenge in both image processing and pattern recognition. There are different approaches to achieve higher performance for this problem, based on support vector machines, neural networks and nearest neighbor methods.

While solving handwritten digits recognition we come across many challenges, one being the digits written are not always the same, they may vary in size and thickness. The variety in handwriting influences the appearance and formation of digits. Our goal is to recognize the handwritten digits (0-9) from the dataset of images.

III. Background

Recognition of handwritten digits is a big challenge since 1980's especially in the old manuscripts and documents. Digit recognition with use of classifier gives a greater performance and use such as recognition of zip codes. The most general problem in digit recognition is to predict the similar digits like (1, 7), (5, 6), (3, 8). Also, users write same digits in different ways. This influences the appearance and shaping of digits. We can perform the same experiment by implementing different kernel functions such as polynomial kernel function, radial basis function and sigmoid function.

Numerous experiments are also being conducted in high dimensionality and small sample size to analyze the superior classification performance of SVM.

IV. METHODOLOGY

1. We will analyze the dataset by implementing learning techniques like Support Vector Machines (SVM) to know the accuracy of the datasets.
2. The data is pre-processed and image is formatted/re-shaped.
3. The trained data is stored and tested against the testing data.
4. The approach to recognize digits involves the following steps
    - Develop an algorithm
    - Train
    - Predict
    - Save the handwritten digits.

SVM is known for excellent generalization in machine learning problems. SVM generates a hyperplane which can be used as a classification plane and SVM can also be used for regression.

V. EXPERIMENTS

Load the data:

We are getting data from kaggle and trying to load data, which is like, Training set has 42000 rows and 785 columns and Test set has 28000 rows and 784 columns.

The database:

The train.csv has 42k rows, where the first column is digit labels and remaining 784 columns are pixel color values that are from 0 to 255 and the test.csv contains 28k rows of pixel color values which are classified as digits. The training set has shape of (42k, 784), each row is 28*28 pixels image. Which will be reshaped to have the training set of each row 28*28 matrix of pixel color values and same is done with testing dataset.

Train SVM classifier:

Here we do ‘fit’ between training set of images and test set of what those images mean. We now try to predict the validation data using score, which tells how accurate our model is, like what percentage of getting our hand writing correct. For example, if score = 0.8312, our algorithm will get 83 images right for every 100 images in the dataset.

VI. ANALYSIS

We have loaded the csv files as input data and obtained the number of rows and columns from the files. We are reshaping the pixel size of the images and predicting the accuracy by training the dataset. Now, further we hope to perform dimensionality reduction with the help of PCA and keras for image augmentation for predicting new dataset and save that to .csv file.

VII. CONCLUSION:

This project work deals with recognition of handwritten digits by applying Support Vector Machine technique and data processing. At the end of this experiment we will determine the accuracy rate that the current feature extraction and classification techniques can achieve in handwritten digit recognition on a known database.

VIII. REFERENCES

[1] Digit Recognition. (n.d.). Retrieved November 03, 2016, from http://rahular.com/digit-recognition/

[2] MNIST database. (n.d.). Retrieved October 24, 2016, from https://en.wikipedia.org/wiki/MNIST_database

[3] Handwritten Digit Recognition using Convolutional Neural Networks in Python with Keras - Machine Learning Mastery. (n.d.). Retrieved November 01, 2016, from http://machinelearningmastery.com/handwritten-digit-recognition-using-convolutional-neural-networks-python-keras/

[4] (n.d.). Retrieved November 04, 2016, from https://www.kaggle.com/fgrzeszc/digit-recognizer/multiclass-svm-with-artificial-training-samples/run/410615

[5] Handwriting Recognition with Python. (n.d.). Retrieved November 04, 2016, from http://www.youtube.com/watch?v=PO4hePKWIGQ