Handwritten digit recognition using KNN

Gayathri Rajkumar

Grand Canyon University: DSC 540

09/16/2020

This technical report presents the method of analyzing handwritten digit recognition using the k-nearest neighbor algorithm.

**Introduction**

Recognizing the handwritten character is one the import and critical tasks as handwritten signatures are used in various applications includes bank loan process, property purchase, postal, etc. Some handwritten letters or documents can help to identify the criminal cases. In offline handwriting recognition, the text is analyzed after being written. This paper presents how the offline handwritten digits are getting classified.

**Data and Preprocessing**

**Data**

The MNIST (Modified NIST) database contains offline handwritten digits comprise of images of digits from 0 through 9 with their corresponding labels. It contains 60,000 training images and 10,000 test images. This is a subset of NIST database. The MNIST has sized-normalized images and centered in a fixed-sized image. The MNIST database is subset of NIST, it was created from NIST’s special database 3(SD3) and 1 which contains binary images of offline handwritten digits. SD3 is much easier and cleaner than SD1. The MNIST database created using 50% Data from SD3 and 50% SD1, 30000 training data is from SD3 and 30000 training data from SD1.

**Preprocessing the binary image**

Data preprocessing is an important task to achieve high performance. In this stage, offline handwritten digit database images are getting converted to binary images and remove extract black dots or lines. It is necessary to remove noise from the background image to achieve high performance. The algorithm uses 3 X 3 templates to identify and remove the noise. It assumes assumed that the pixel (p, q) as center pixel and neighbors of the point (p, q) are (p- 1, q), (p-1, q+1), (p, q+1), (p+1, q+1), (p+1, q), (p+1,q-1), (p, q-1), and (p-1, q-1) (Ravi Babu et. al, 2014).

**Feature Extraction method:**

Feature extraction is a crucial task in this process in the image recognition process. Identifying the stroke of the digit is important to classify the digit. Digits are written by different users and their handwriting is different from one to another. In this phase, algorithms identify the strokes and patterns from the training dataset. The k-nearest neighbor(k-nn) is a supervised learning algorithm that helps to classify the images.

**Classification**

The k-NN is a classification algorithm that is one of the recommended algorithms for image recognition. The k-NN uses Euclidean distance to find the distance between two data points and using that it determines the nearest neighbors based on the pairwise distances.

**Euclidean distance**

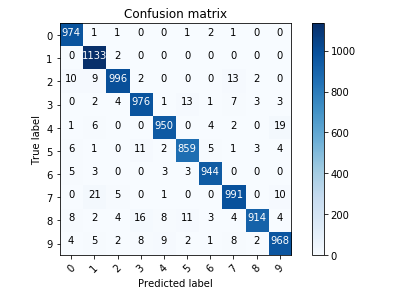
One of the popular methods for measuring the distance between two patterns is Euclidean distance. Euclidean distance is not a good choice if there is a strong correlation between data points. It is also sensitive to outliers. It depends highly on scale, and variables that possess bigger scales impact the total distance to a great extent. Thus, we first normalize continuous measurements and only then calculate the Euclidean distance. This transforms all measurements to the same scale (Gopal, 2019). The k-nn algorithm uses minimum distance to classify the images into under correct label.



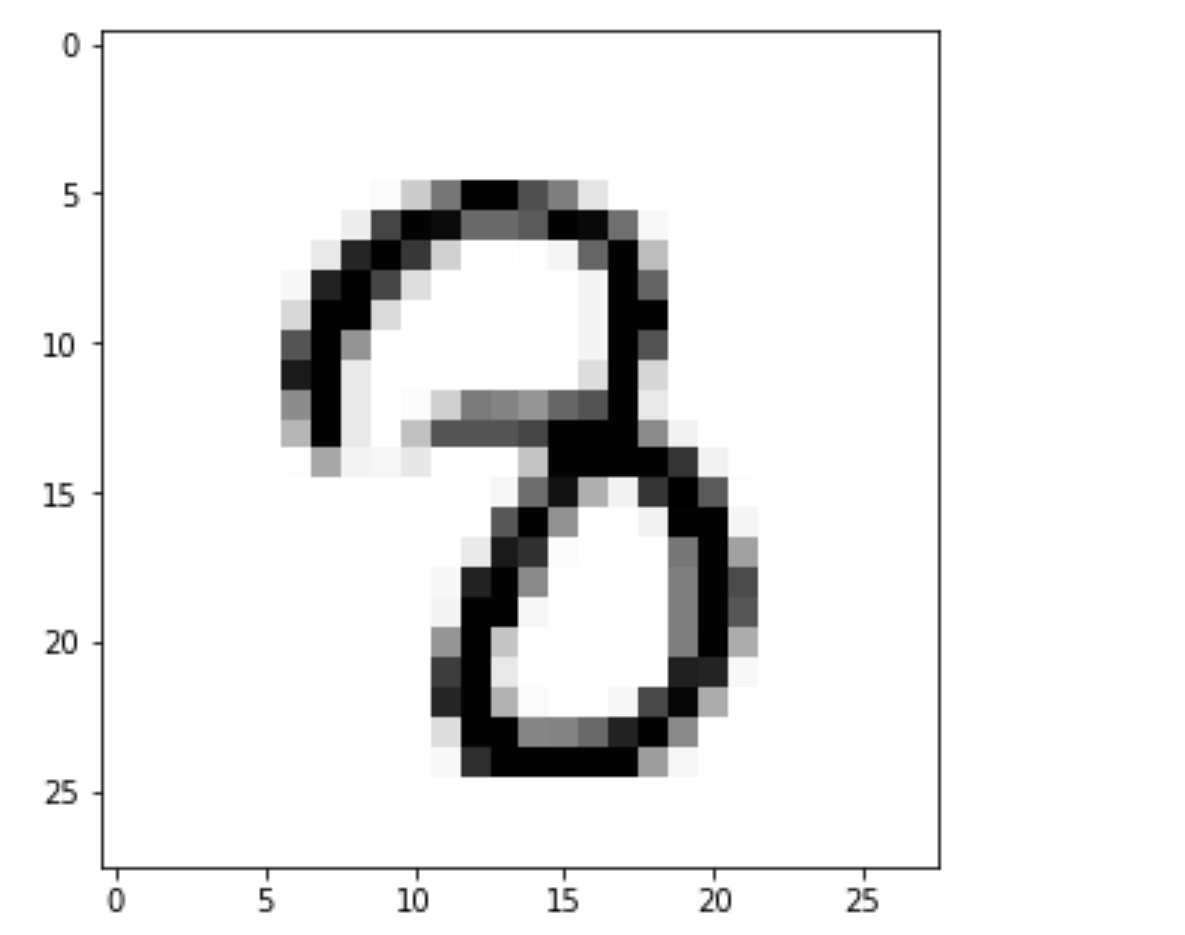
**Result**

This algorithm executed using different k values for 10,000 test datasets, and the overall accuracy of the model is 98%. The digit 9 has an accuracy of 95%, digits 0, 1 have 99%, digit 2 -5, 7 have the 96%, digit 6 has 98%, and digit has 93%.

|  |  |
| --- | --- |
| K value | Accuracy |
| 1 | 0.9847 |
| 2 | 0.9801 |
| 3 | 0.9874 |
| 4 | 0.9854 |
| 5 | 0.9867 |
| 6 | 0.9847 |



Some of the digit are misclassification error in this model, for example, there four 8 digits classified as 2. The below figure 8 is classified as 2, because the loop in digit 8 is not closed properly. The Overall error rate is 2%, and an accuracy is 98%. There are 9705 digits classified as true positive and 295 are false positives. In ROC curve, 295 digits fall under the triangle and bottom right of x-axis. The digit 7 has majority vote.



References

Gopal, M. (2019). *Applied machine learning*. New York: McGraw-Hill Education

Ravi Babu, U., Chintha, A. K., & Venkateswarlu, Y. (2014). Handwritten Digit Recognition Using Structural, Statistical Features and K-nearest Neighbor Classifier. *International Journal of Information Engineering and Electronic Business*, *6*(1), 62–68. https://doi-org.lopes.idm.oclc.org/.DOI: 10.5815/ijieeb.2014.01.07.

The MNIST DATABASE. (n.d.). Retrieved from <http://yann.lecun.com/exdb/mnist/>