Handwritten digits recognition with decision tree classification: a machine learning approach

Tsehay Admassu Assegie, Pramod Sekharan Nair
Department of Computing Technology, College of Engineering and Technology, Aksum University, Ethiopia

Article Info

Article history:

Received Jan 29, 2019
Revised Apr 22, 2019
Accepted May 4, 2019

Keywords:

Decision tree classification
Handwritten digits
Handwritten digits recognition
Machine learning

ABSTRACT

Handwritten digits recognition is an area of machine learning, in which a machine is trained to identify handwritten digits. One method of achieving this is with decision tree classification model. A decision tree classification is a machine learning approach that uses the predefined labels from the past known sets to determine or predict the classes of the future data sets where the class labels are unknown. In this paper we have used the standard kaggle digits dataset for recognition of handwritten digits using a decision tree classification approach. And we have evaluated the accuracy of the model against each digit from 0 to 9.

Corresponding Author:

Tsehay Admassu Assegie,
Department of Computing Technology,
Aksum University,
POB, 1010, Ethiopia
Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

Handwritten digits recognition is the process of identifying digits written by human using a machine learning approach [1, 2, 3, 4, 5, 6, 7, 8, 9]. Machine that performs the recognition task has to be trained so that the machine can perform the task based on the training given with the past data sets with the known class labels. Machine learning models such as, artificial neural network, convolutional neural network, decision tree classifier and improved chain code histogram feature [1, 2, 3, 4] can also be applied to handwritten digits recognition problem, however, the recognition of handwritten digits is still a research issue, numerous researches have been carried out and none of them has proposed a perfect approach to recognize handwritten digits. This implies that the problem of recognizing a handwritten digit is an important research issue and there are different approaches and algorithms proposed by researchers for recognitions of handwritten digits.

The main problems of handwritten digits recognition using machine learning approaches are listed as follows [9]:

1) Handwriting styles vary based on the writer- it is usually very difficult even for human to recognize handwritten digits because of the significant difference of handwriting styles.
2) Similarity between handwritten digits, for example six and four may look the same digit based on the handwriting style of the writer similarly, one and seven may look the same digit.
3) There is no perfect machine learning model that is suitable and effective in recognizing handwritten digits. Different researchers use different models but no model is perfectly suited for recognition of handwritten digits. This issue is focused in this paper. Efficiency and accuracy of proposed decision tree classification model is evaluated.
4) The handwritten digits cannot be recognized by using fixed image recognition patterns, because they are different for different writers.

The standard data sets like kaggle and MINST have simplified the handwritten digits recognition problem [10, 11]. Those datasets contains large number of sample images of handwritten digits which can be used in machine learning models like decision tree classification without preprocessing and the effort of preparing dataset is avoided. But still some of the questions remained unanswered and will continue to exist in the feature, as there is no perfect model used in machine learning in recognition of handwritten digits.The problems addressed in this paper are, how effective a decision tree classification in handwritten digits recognition. And the implementation and accuracy of decision tree classification in handwritten digits recognition is evaluated using python programming language.

Decision tree classification

A decision tree classifier is a machine learning model widely used to solve classification problems, like handwritten digits recognition. As the name decision tree suggests, the model breaking down our data by making decisions based on asking a series of questions. Let's consider the following example where we use a decision tree to decide upon recognition of a 0 and a 1 digit. The classifier tests all of the possible sets with digits 0 and 1 until it finds a matching digit as shown in Figure 1.

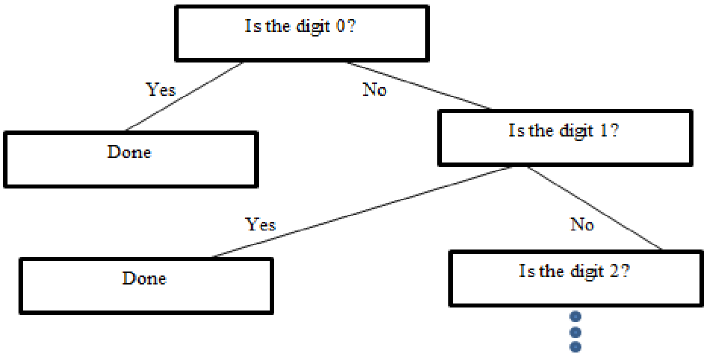
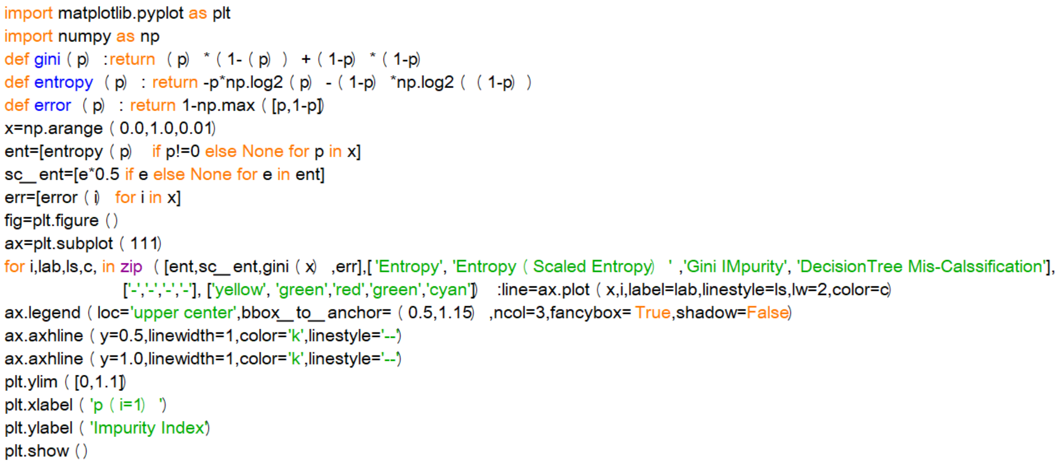


Figure 1. Decision tree classification

The success rate of decision tree classifier in python

To test the success rate of a decision tree classification in recognition of handwritten digits we have used python programming, a very powerful language for implementation of machine learning algorithms, many of the researchers are depending on python language for machine learning algorithm implementations. The code shown below is used to draw the graph for success rate of the classifier, shown in Figure 2 and the entropy, the impurity for the decision tree classification is shown in Figure 3.



```
import matplotlib.pyplot as plt
import numpy as np
def gini ( p) :return ( p) * ( 1- ( p) ) + ( 1-p) * ( 1-p)
def entropy ( p) : return -p*np.log2 ( p) - ( 1-p) *np.log2 ( ( 1-p) )
def error ( p) : return 1-np.max ( [p,1-p])
x=np.arange ( 0.0,1.0,0.01)
ent=[entropy ( p) if p!=0 else None for p in x]
sc_ent=[e*0.5 if e else None for e in ent]
err=[error ( i) for i in x]
fig=plt.figure ()
ax=plt.subplot ( 111)
for i,lab,ls,c, in zip ( [ent,sc_ent,gini ( x) ,err],[ 'Entropy', 'Entropy ( Scaled Entropy) ' ,'Gini IMpurity', 'DecisionTree Mis-Calssification'],
                    ['-','-','-','-'], ['yellow', 'green','red','green','cyan']) :line=ax.plot ( x,i,label=lab,linestyle=ls,lw=2,color=c)
ax.legend ( loc='upper center',bbox_to_anchor= ( 0.5,1.15) ,ncol=3,fancybox= True,shadow=False)
ax.axhline ( y=0.5,linewidth=1,color='k',linestyle='--')
ax.axhline ( y=1.0,linewidth=1,color='k',linestyle='--')
plt.ylim ( [0,1.1])
plt.xlabel ( 'p ( i=1) ')
plt.ylabel ( 'Impurity Index')
plt.show ()
```

Figure 2. Python code for modeling the success rate of decision tree classification

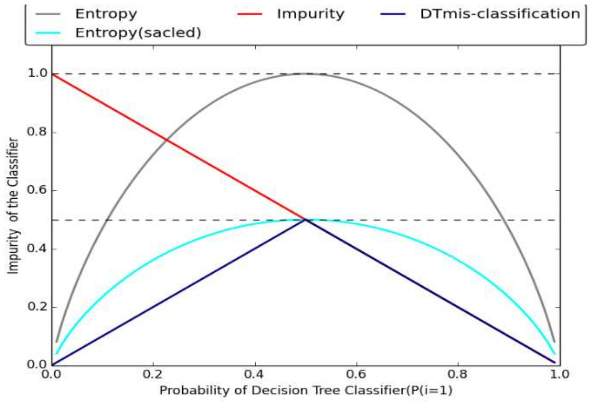


Figure 3. Success rate of decision tree classifier

1. LITERATURE REVIEW

In this section, the previous works related to handwritten digits recognition will be reviewed. The studies are discussed in the following sections. Handwritten digits recognition can be implemented with scanning the digital images pixels and extracting the features using a neural network as classifiers for handwritten digits recognition [1].

Handwritten digits recognition is becoming a greater demand as one of the computer vision [2] techniques. Computer vision has many applications for interfacing human and a machine. The focus of the author was getting better accuracy and performance requirement with convolutional neural networks for classification.

A blob based classification of image is another algorithm used to solve handwritten digits recognition problems. The algorithm recognizes digits by classifying them as blobs with and without stem [3]. In this method the mathematical morphology was used to construct the classification models. This is a technique of recognition of digits by identification of blobs and stems. The problem with this method is that it is not able to recognize the broken digits of large gap and digits with extra stems.

Multiple perception layers with neural network is another model used in handwritten digits recogniton [4]. In their study, the authors used MNIST dataset for handwritten digits recognition, the dataset was trained with gradient descent back-propagation algorithm and the verified with the feed-forward algorithm.

A Decision tree classification is one of the simplest machine learning model used in handwritten digits recognition [5] and a comparative study performed on decision tree classification and random forests. According to authors, the random forest performed superior than decision tree classification. One of the most commonly used machine learning model in handwritten digits identification is KNN (K-Nearest Neighbor) [6]. In this study MINST dataset was used for testing the algorithm, 50,000 images were used for training and 550 images for testing the algorithm. According to authors, the precision level of the system is 70.9 % and the accuracy of the system will improve based on the quality of dataset.

A number of machine learning models, such as KNN, K Nearest classification, Artificial Neural Network, ANN Bayesian Classification and Decision tree classification can also be used to identify handwritten digits as machines learning techniques. The studies that we have performed on accuracy of each of the training models differs, the ANN [7, 12, 14] is better performed although it has its own drawbacks. The accuracy of the machine learning models are also depends on the quality of the image to be predicted and the availability of the inferiority of digit images in the training sets. Computational efficiency is another parameter to select better model and the KNN is better in efficiency compared with other algorithms [13].

2. DATASET AND FEATURE SELECTION

2.1. Kaggle dataset

In this paper, we have used kaggle dataset. The kaggle dataset has sample handwritten digits for evaluating machine learning models on the handwritten digit recognition problem. It has been commonly used in research to implement handwritten digit recognition systems. The kaggle dataset contains 21,000 training and 21,000 testing of handwritten digits (0 to 9). Each digit is standardized and cantered in a gray-scale (0 - 255) image with size 28 × 28 pixel. Each image consists of 784 pixels that represent the structures of the digits. Some of the examples from the kaggle dataset are shown in Figure 4.

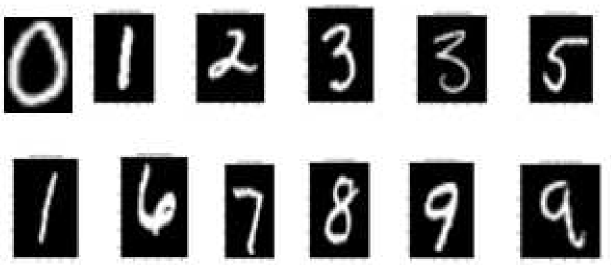
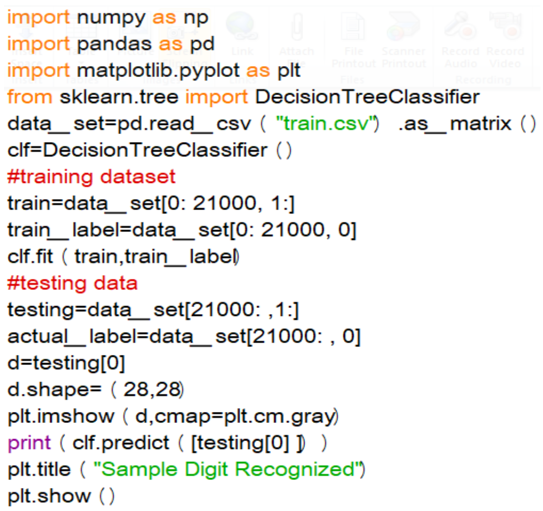


Figure 4. Examples of kaggle dataset

2.2. Digits recognition

Decision tree classification model is used to train 42000 dataset; by splitting, half of the datasets for the training set and the remaining 21000 data sets for testing. We have followed the following steps to perform the classification on kaggle handwritten digits dataset and the pseudo code of the handwritten digits recognition is shown in Figure 5.

a. Load the kaggle handwritten digits datasets for classification.
b. Split the data sets into two sets, one for the training and the rest for testing.
c. Learning and predicting, in the digits datasets, the recognizer is trained to predict, given an image of a handwritten digit. We have used the samples of each of the 10 possible classes of handwritten digits from kaggle dataset (the digits zero to nine) on which we fit an estimator to predict the classes to which unseen samples belong to.
d. Testing the accuracy of the classifier.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
data__set=pd.read__csv ( "train.csv")  .as__matrix ()
clf=DecisionTreeClassifier ()
#training dataset
train=data__set[0: 21000, 1:]
train__label=data__set[0: 21000, 0]
clf.fit ( train,train__label)
#testing data
testing=data__set[21000: ,1:]
actual__label=data__set[21000: , 0]
d=testing[0]
d.shape= ( 28,28)
plt.imshow ( d,cmap=plt.cm.gray)
print ( clf.predict ( [testing[0] ])  )
plt.title ( "Sample Digit Recognized")
plt.show ()
```

Figure 5. Pseudo code of digits recognition

3. EXPERIMENTAL RESULTS

The experiment was conducted on handwritten digits of the standard kaggle dataset using the decision tree classifier for training the machine. In this dataset images were preprocessed so in this paper preprocessing was not included. This dataset contains 42,000 images of sample handwritten digits. Form these images; we have used 21,000 in training and the rest 21,000 images in testing. We tested the decision tree classification model on all of the digits (0-9). Table 1, shows the test results after training the machine on this dataset. In Table 1, it is shown that some of the digits are not recognized by the machine learning model, the decision tree classifier. For example 8 were predicted as 5 and the reason behind this is the similarity of handwriting style between 5 and 8. In Table 1, the sample handwritten digits recognized by the decision tree classifier is shown and the accuracy of the classifier in recognizing each digit (0 to 9) is shown in Table 2.

Table 1. The worest recognized digits lists

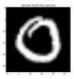
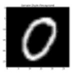
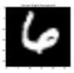
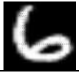
Handwritten Digits images/Test Data	System Digit Prediction	Actual Expected Prediction
	Correctly Predicted	0
	Correctly Predicted	0
	Correctly Predicted	0
	Wrongly Predicted	0
	Wrongly Predicted	6
	Correctly Predicted	6
	Correctly Predicted	6
	Correctly Predicted	6

Table 2. Accuracy of the decision tree classifier in handwritten digits recognition

HandWritten Digits(0-9)	Accuracy of the Classifer (100%)
0	83.56
1	93.73
2	83.69
3	83.73
4	83.81
5	83.65
6	83.47
7	83.81
8	84.12
9	83.75

Figure 6, shows the accuracy of the decision tree classifier (in percentage) in recognizing each handwritten digit. And the digit 1 is recognized with better accuracy by the classifier, 8 is recognized with least accuracy and this is because the digit 8 resembles with many other digits for example depending on the hand writing style, it may look five. This shows that, the difference in handwriting style have a great influence on the accuracy of a machine learning models used in handwritten digits recognition.

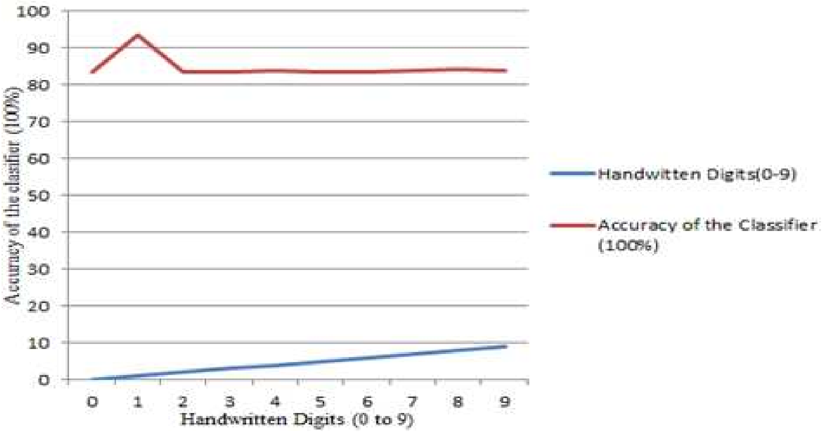


Figure 6. The accuracy of decision tree classifier

4. CONCLUSION

In this paper, we have used a decision tree classification model for machine learning that recognizes handwritten digits. The machine was trained with a kaggle dataset which contains 42000 rows and 720 columns and result shows 83.4% accuracy. We have used digit images pixels as features vector and Decision Tree as classifiers for handwritten digits recognition. We have used the Kaggle repository for training and testing datasets, the experiment result shows that the decision tree classifier is effective in recognition of handwritten digits.

REFERENCES

[1] Kh Tohidul Islam, Ghulam Mujtaba, "*Handwritten Digits Recognition with Artificial Neural Network*," *Proc. of the International Conference on Engineering Technologies and Technopreneurship (ICE2T 2017),* Sep 2017.

[2] Haider A. Alwzwazy, Hayder M. Albehadili, Younes S. Alwan, Naz E. Islam. "Handwritten Digit Recognition Using Convolutional Neural Networks," *International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization),* vol. 4, no. 2, Feb 2016.

[3] Vijaya Kumar, A Sirikirshna, Raveendra Babu and Radhika Mani, "Classification and recognition of handwritten digits by using mathematical morphology," *Indian Academy of Sciences*, 2010

[4] Ismail M. Keshta. "Handwritten Digit Recognition based on Output Independent Multi-Layer Perceptron's," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 8, no. 6, 2017.

[5] Lavanya K, Shaurya Bajaj, Prateek Tank Shashwat Jain, "Handwritten Digit Recognition using Hoeffding tree, Decision tree and Random forests Comparative Approach," *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)*,2017.

[6] Nurul Ilmi, Tjokorda Agung Budi W, Kurniawan Nur R, "*Handwriting Digit Recognition using Local Binary Pattern Variance and K-Nearest Neighbor Classification*," *IEEE, Fourth International Conference on Information and Communication Technologies (ICoICT)*, 2016

[7] Stuti Asthana1, Amitkant Pandit, Dinesh Goya, "A Literature Survey On Better System Efficiency Of Handwritten Numeral Recognition," *Suresh Gyan Vihar University International Journal of Environment, Science and Technology,* vol. 3, no. 1, pp. 1-5, Jan 2017.

[8] Shengfeng Chen, Rabia Almamlook, Yuwen Gu, Dr. Lee wells, "*Offline Handwritten Digits Recognition Using Machine learning*," *Proceedings of the International Conference on Industrial Engineering and Operations Management Washington DC*, USA, Sep 27-29, 2018.

[9] Anuj Dutt, AashiDutt, "Handwritten Digit Recognition Using Deep Learning," *International Journal of Advanced Research in Computer Engineering & Technology,* vol. 6, no. 7, Jul 2017.

[10] Pooja Yadav, Nidhika Yadav, "Handwriting Recognition System- A Review," *International Journal of Computer Applications,* vol. 114, no. 19, pp. 46-40, Mar 2015.

[11] Areej Alsaafin, Ashraf Elnagar, "A Minimal Subset of Features Using Feature Selection for Handwritten Digit Recognition," *Journal of Intelligent Learning Systems and Applications*, vol. 9, pp. 55-68, 2017.

[12] J. J.Wang, S.G. Hu1, X.T. Zhan, Q.Yu1, Z. Liu, T. P. Chen, Y.Yin, Sumio Hosaka, Y. Liu, "Handwritten-Digit Recognition by Hybrid Convolutional Neural Network based on HfO2 Memristive Spiking-Neuron," *Sceintifi Reports*, Aug 2018

[13] Stuti Asthana, Amitkant Pandit, Dinesh Goyal, "A Literature Survey On Better System Efficiency Of Handwritten Numeral Recognition," *Suresh Gyan Vihar University International Journal of Environment, Science and Technology,* vol. 3, no. 1, Jan 2017

[14] Raul R. Tiwari, Aparnavis Hwana, Dhanashree Wadhone, "Handwritten digits recognition using back propagation nueral network and KNN classifier," *International Journal of Electrical, Electronics and Data Communication*, vol. 1, 2013.