COMPARATIVE STUDY OF VARIOUS MACHINE LEARNING ALGORITHMS IN HANDWRITTEN CHARACTER RECOGNITION

Vipin Garg*, Shubham Gupta*

*B.Tech Student, Maharaja Agrasen Institute of Technology, Rohini, New Delhi

Dr. Mrs. Bhoomi Gupta[1]

1 Assistant Professor, Maharaja Agrasen Institute Of Technology, Rohini, New Delhi

ABSTRACT

The HCR has been an interesting area of research and has numerous applications like getting information from data entry, cheque, application for loans. handwritten characters are difficult to recognize due to various writing styles and difference in size and shape of letters. In this paper handwritten characters are recognised using various supervised machine learning methods like logistic regression, random forest, k -nearest neighbour (KNN) and Convolution neural networks (CNN) which are used to get better recognition rates. We then considered the accuracy and efficiency of these strategies independently and as a whole.

Keywords: handwritten recognition, machine learning, supervised learning, logistic regression, random forest, k -nearest neighbour (KNN), SVM and Convolution neural networks (CNN).

INTRODUCTION

Character recognition is a field in which lot of work has been done but not much for analysing a complete document. there are many applications of recognising the text in a document like reading medical prescriptions, bank cheques and other official documents. It can also be used in detective or police departments in applications like handwriting based person identification, identifying real from forged documents, etc. [14]

OCR is the process in which printed or written text characters are recognised by a computer. This is done by photo scanning of the text character-by-character, analysing the scanned-in image, and then translating the character image into character codes, such as ASCII, commonly used in data processing.[15]

In OCR processing, light and dark areas of the scanned-in image or bitmap are analysed in order to identify each alphabetic letter or numeric digit. When a character is recognized, it is converted into an ASCII code. To speed up the recognition process Special circuit boards and computer chips designed expressly for OCR are used.

Libraries use OCR to digitize and preserve their holdings. the OCR can also be used to process checks and credit card slips and sort the mail. Billions of magazines and letters are sorted every day by OCR machines, considerably speeding up mail delivery.[12]

Optical character recognition is of two types:1) Online Character Recognition 2) Offline character recognition.[13]
1) Online Character Recognition: In online character recognition character are recognised at a time of writing using a device like electronic pen in which character are recognised by analysing pen movement.
2)Offline character recognition: In offline character recognition character images are scanned first and then taken as input for recognition.

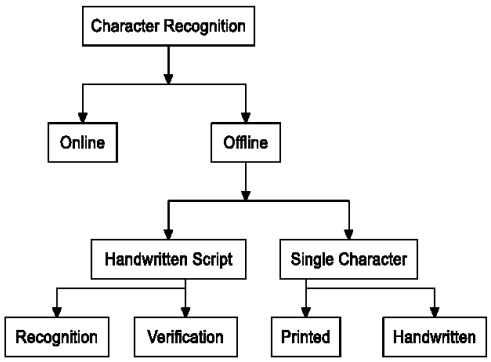


Fig. 1 Types of character recognition

Image source: https://www.semanticscholar.org/paper/Optical-Character-Recognition-using-Neural-Network-Varshney-Chaurasiya/05fbdf4caeb948afa2fb11c31c51447c4d82bc06

to acquire information from customers most of the organisations use documents. These documents are generally handwritten. Such documents can be forms, checks, etc. documents are transformed and stored in digital formats for easier retrieval and information collection. Common practice to handle that information is manually filling same data into computer. It would be tiresome and time consuming to handle such documents manually. Hence a special Handwritten Character Recognition Software is required which will automatically recognize texts from image of documents. Handwritten Character Recognition (HCR) Software had made the process of extracting data from documents and converting into digital formats easier. Banking sectors, Health care industries and many such organizations where handwritten documents are used regularly. HCR can be useful in newly emerging areas where handwriting data entry is required, such as development of electronic libraries, multimedia database etc.

MOTIVATION

In field of character recognition lots of works is done but there are less cases where complete documents are analysed. Character recognition has many useful application like recognising text in medical prescription, bank cheques and other documents. Also used in application like handwriting based on person identification identifying real from forged documents and in various police and detective departments. Handwriting recognition can be broken into a number of relatively independent modules.

After reading lots of articles and research papers on handwritten character recognition we have applied various strategies and structures for these modules. We have considered the efficiency of these module as a whole and independently also. We have used the directional graph for the feature extraction and relative position matching The aim of our project is to implement these strategies to get efficient accurate and scalable handwriting recognition software

DATASET

MNIST ("Modified National Institute of Standards and Technology") is best open source to get the dataset of computer vision. MNSIT is reliable source where researchers and students can get access to data for testing machine learning algorithms. We have used MNSIT hand written data for our research task.In our dataset the digits are taken from various scanned document's which are then normalized in size and centred. The data files contain grey-scale images of hand-drawn digits, from zero through nine. Each image has 28 pixels in width and 28 pixels in height, which is 784 pixels in total. . Each pixel has a single pixel-value associated with it, indicating the darkness or lightness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.[1]

Image source: http://neuralnetworksanddeeplearning.com/chap1.html

WORKING PRINCIPLE

It is known that handwritten character comprises of six phases which include image acquisition, pre-processing and segmentation as the primary considerations, and the rest are included in the diagram as shown in Figure 1.

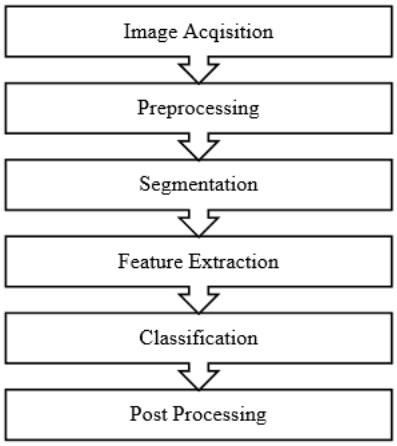


Figure 1: Block Diagram of Character Recognition

Image source: Ayush Purohit et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, 1-5

1)Image Acquisition: It is very clear through the linguistics of the words involved, yet if it is required to state, Image Acquisition is an action of retrieving an image from any given source. This is usually a hardware-based source for processing. Conclusively, it can be thought as the first step in the sequence. This is because otherwise, no processing would be possible without an image.[11]

2)Pre-processing: Because of the fact that a given data has to be pre-processed after it being selected, one has to follow the given steps:
One has to format the given data. This has to be done to make it good for machine learning.
The data has to be cleaned to remove redundancy.
The data has to be sampled. Again, this is to reduce algorithm's running times and other constraints.
Conclusively, data has to be filtered, and thereby cleaned, on the basis of the provided variables:

Insufficient Data
Now, machine learning algorithms requires a large amount of data. For this, we need thousands, or sometimes even millions, of sample data instances. This quantity of these instances depends directly upon how complex the problem is, let alone the chosen algorithm.[12]

Non-Representative Data
It is required for the samples selected to be exact representations of the presented data, since otherwise the data might train the desired algorithm, such that it won't work well on the kept-aside test data.[12]

Substandard Data
As clearly definitive, the possible errors, outliers and obtained noise can be nullified so as to obtain a better fitting model for the algorithm. In addition, the average value of attributes could be fully ignored.
One example is the age attribute: the age of about ten percent of the audience, say, could be ignored.[12]
During data preparation, it becomes important to select the correct size of the sample. This is because otherwise one could get skewed results for the samples too large or too small.[12]

Sampling Noise
Small samples get trained on some non-representative information, that causes sampling noise. Checking the sentiments of people is an example.[12]

Sampling Bias
For large samples there is no sampling bias, which makes them work well, therefore the right information is selected. For instance, there definitely would be some sampling bias that might occur while checking people's sentiment for only the technical sound set of voters, all the while ignoring others. [12]

Normalization: As is definitive, normalization is the technique to scale the sole samples in order to own a unit norm. This can be beneficial in case you plan on using some quadratic form like the dot-product or even some other kernel for quantifying the similarity of any defined pair of samples.

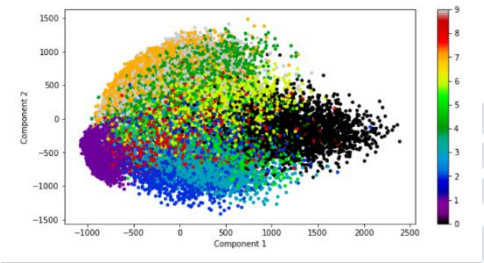
3)Segmentation:

It is the technique to decompose an image into sub-images consisting of individual characters. It comprises of Line, Word and Character Segmentation. In line segmentation a given line is separated from its parent paragraph. In word segmentation a given word is separated its parent line. In character segmentation Character is separated from its parent word.

4) Feature extraction:

1. Image pixel vector – Image pixel vector is the simplest method used to extract the image features. In image pixel vector to train learning models we have used the intensity of image pixels as the feature vector. For processing of image we have used the greyscale version of the image. As compared to other feature extraction methods this method is fast but has little low accuracy models for feature extraction.

2.PCA: The dimensionality of a data set consisting of many variables corelated to each other is reduced with the help of principal component analysis (PCA) while variation in dataset is retained upto maximum extent during this process. The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. The dataset must be scaled that is used for PCA technique. The result obtained by sing PCA technique are sensitive to the relative scaling. In terms of layman language it can be stated as method of summarizing the data.[17]



It is interesting to see ho w well PCA separates the feature space into visible clusters already for 2 components. In our project we have used 324 components to capture 99% variance in the data.

5)Classifications:

Classification is the most appreciated way for making decisions in the domain of recognising characters.
The classifier inputs the older features that were extracted. Now, the distinct classifier types considered in the file are LR, RF, SVM, KNN and CNN.

Logistic regression: It is one of the most reputed algorithms in ML field, of course after not considering linear regression. In most ways, linear regression and logistic regression are pretty much related in terms of concept. The disparity in their use in the field; Linear regression algorithms are preferred in prediction purposes while logistic regression is preferred in classification.[18]
Cost Function for logistic regression is:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Now this is done with a sigmoid function:

$$h = g(z) = \frac{1}{1 + e^{-z}}$$

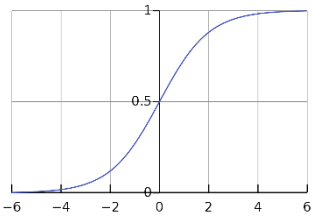


Image source: https://en.wikipedia.org/wiki/Sigmoid_function

Random forest: You can think of these as some extensions of decision trees that include the averaging different decision trees in order to produce better results. Because of low variance than DTs, plus the fact that it doesn't show any overfitting on increasing the count of classifiers, these are correctly preferred.

KNN: Similarly, KNN concerns the closeness of sample features' resemblance to the training set to estimate the way we differentiate the way a given data point is classified. Based on the votes of neighbours, a given object is classified. This is done by assigning the object to the most common class to the k-nearest neighbours.We consider that our data set has been represented as a matrix $D = N \times P$, that contains P cases $s^1 \ldots s^P$, where each case s^i has N features $s^i = \{s_i \ldots s^i\}$. A vector o with length P of output values $o = \{o_1 \ldots o^P\}$ accompanies this matrix, listing the output value o^i for each scenario s^i.
KNN algorithm follows these steps:
Output values of the M nearest neighbours to query scenario q is stored in vector $r = \{r_1 \ldots r^M\}$
Arithmetic mean output across r as follows:

$$\bar{r} = \frac{1}{M} \sum_{i=1}^{M} r_i$$

is the output value for the query scenario q.[5]

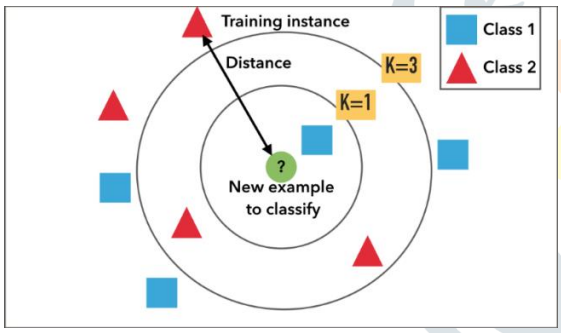


Image source: https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

SVM Algorithm: An SVM is a discriminative classifier that was officially defined by a separating hyperplane. In the layman language, for a given labelled training data, the SVM algorithm gives an outcome hyperplane that can be used to categorize new examples.Decision planes was the inspiration in the development of the SVMs, that defined decision boundaries. The main purpose of a decision plane is to separate a set of objects that have different class memberships.[19]

Convolutional Neural Networks: A neural network is a machine learning model which consists of connected layers of *neurons*. A neuron contains a number, the so-called *activation*. Connections are assigned *weights*, which describes the strength of the signal to the connected neuron. Input data is fed into the first layer, activating each input neuron to some extent.The network uses the weights and activation function to determines which neurons from the next layer to activate and the strength of activation based .This is *feedforward* process is continued until the output neurons are activated. The architecture of a neural network has a huge influence on which data it can work with and its performance. The following figure illustrates a simple neural network with three layers.

CNNs are the neural network of special types. CNNs can be classified into two parts : A *feature learning* part and a *classification* part. Each part consists of one or multiple layers. Feature learning is genrally performed by combining two types of layers: *Convolution layers* and *pooling* layers. Classification is then performed based on the learned features through *dense layers*, also known as fully connected layers. Additionally, there is an *input layer*, containing the image data, as well as an *output layer*, containing the different classes we are trying to predict.

The following figure illustrates a CNN with one convolution layer, one pooling layer, and one dense layer. The task is to predict whether the image depicts a cat. Layers that are in-between the input and output layer are also referred to as *hidden layers* as there state is not directly visible when treating the model as a black box.

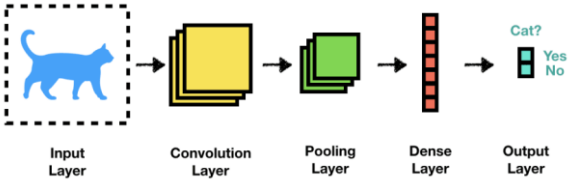


Image Source: https://dev.to/frosnerd/handwritten-digit-recognition-using-convolutional-neural-networks-11g0

The input layer consists of 784 units and our neural network consists of 2 hidden layers with 512 and 256 units respectively. We have used ReLu activation function in both the hidden layers. The output of output layer is normalized using softmax function so that the output represents probabilities.
Softmax

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

After applying softmax function, the cost is calculated using cross entropy.

$$J(Y^T, Y^P) = -1/N(\sum_{i}^{N} \sum_{K}^{classes} Y^T log(Y^P) + (1 - Y^T) log(1 - Y^P))$$

The model is initialized with random weights and using cost function we update the weights using back-propagation. We have used batch size of 128 for determination of cost function. We have run 12 epocs to converge the cost function

CONCLUSION:

We obtain the result as follows:
Accuracy of Logistic Regression Algorithm is 0.9126984126984127
Accuracy of Random Forest Algorithm is 0.9373809523809524
Accuracy of KNN Algorithm is 0.7096825396825397
Accuracy of SVM is 0.9538095238095238
Accuracy of CNN is 0.9903.

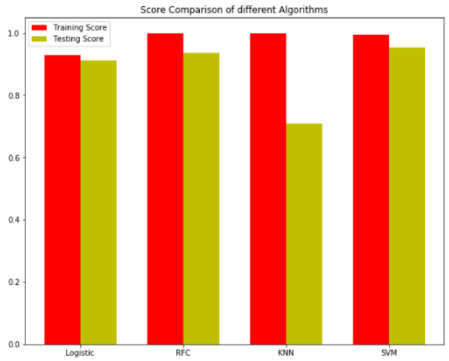


Image: Score comparison of different algorithm used (Logistic, RFC, KNN, SVM)

REFERENCES:

[1] THE MNIST DATABASE OF HANDWRITTEN DIGIT IMAGES FOR MACHINE LEARNING RESEARCH

[2] Applied Logistic Regression, 3rd Edition by David W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant.

[3]Gerard Biau , Analysis of a Random Forests Model, Journal of Machine Learning Research 13 (2012) 1063-1095

[4] Machine Learning with Random Forests and Decision Trees: A Visual Guide for Beginners Kindle Edition by Scott Hartshorn

[5] Paul Lammertsma, K-nearest-neighbor algorithm, #0305235

[6] K Nearest Neighbour's - Classification
https://www.saedsayad.com/k_nearest_neighbors.htm

[7] Proceedings of NTCIR-8 Workshop Meeting, June 15–18, 2010, Tokyo, Japan

[8] S B Imandoust et al., Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background , Int. Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610

[9] A Quick Introduction to K-Nearest Neighbors Algorithm
https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

[10] Keiron O'Shea 1 and Ryan Nash, An Introduction to Convolutional Neural Networks, Published in ArXiv 2015.

[11] Ayush Purohit et al, A Literature Survey on Handwritten Character Recognition,(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, 1-5

[12]Bhoomi Gupta, a novel approach for multi exposure image fusion using deep learning ,
http://jardcs.org/backissues/abstract.php?archiveid=6346&action=fulltext&uri=/backissues/abstract.php?archiveid=6346

[13] J.Pradeep , E.Srinivasan and S.Himavathi, DIAGONAL BASED FEATURE EXTRACTION FOR HANDWRITTEN ALPHABETS RECOGNITION SYSTEM USING NEURAL NETWORK, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011

[14] Asha Tarachandani and Pooja Nath ,Address block recognition and beautification software.IIT Kanpur : April 2000,Artificial Intelligence ME 768 Jan-Apr 2000

[15] OCR (optical character recognition)
https://searchcontentmanagement.techtarget.com/definition/OCR-optical-character-recognition

[16] Advances in Vision Computing: An International Journal (AVC) Vol.3, No.1, March 2016

[17] Principal Component Analysis Tutorial
https://www.dezyre.com/data-science-in-python-tutorial/principal-component-analysis-tutorial

[18] Introduction to Machine Learning Algorithms: Logistic Regression
https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36

[19] SVM (Support Vector Machine)—Theory
https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[20] S M Shamim, Mohammad Badrul Alam Miah, Angona Sarker, Masud Rana & Abdullah Al Jobair , Handwritten Digit Recognition using Machine Learning Algorithms ,Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence Volume 18 Issue 1 Version 1.0 Year 2018.

[21] Surya Nath R S*, Afseena S** ,Handwritten Character Recognition – A Review ,International Journal of Scientific and Research Publications, Volume 5, Issue 3, March 2015 1 ISSN 2250-3153