A Novel Method For Handwritten Digit Recognition System

Abstract :

Character recognition plays an important role in themodern world. It can solve more complex problems and makeshumans' job easier. An example is handwritten character recognition. This is a system widely used in the world to recognize zip code or postal code for mail sorting. There are different techniques that can be used to recognize handwritten characters. Two techniques esearched in this paper are Pattern Recognition and Artificial Neural Network (ANN). Both techniques are defined and different methods for each technique is also discussed. Bayesian Decision theory, Nearest Neighbor rule, and Linear Classification or Discrimination is types of methods for Pattern Recognition. Shape recognition, Chinese Character and Handwritten Digit ecognition uses Neural Network to recognize them. Neural Network is used to train and identify written digits. After training and testing, the accuracy rate reached 99%.This accuracy rate is very high.Keywords: Pattern recognition, neural network, handwritten characters.

1. INTRODUCTION AND HISTORY

Character recognition is becoming more and more important in the modern world. It helps humans ease their jobs and solve more complex problems. An example is handwritten

character recognition [4] which is widely used in the world. This system is developed for zipcode or postal code recognition that can be employed in mail sorting. This can

help humans to sort mails with postal codes that are difficult to identify. For more than thirty years, researchers have been working on handwriting recognition. Over the past few years,

the number of companies involved in research on handwriting recognition [4] has continually increased. The advance of handwriting processing results from a combination of various

elements, for example: improvements in the recognition rates, the use of complex systems to integrate various kinds of information, and new technologies such as high quality high

speed scanners and cheaper and more powerful CPUs. Some handwriting recognition system allows us to input our handwriting into the system. This can be done either by controlling a mouse or using a thirdparty drawing tablet. The input can be converted into typed text or can be left as an "ink object" in our own handwriting. We can also enter the text

we would like the system to recognize into any Microsoft Office program file by typing. We can do this by typing 1s and 0s. This works as a Boolean variable. Handwriting recognition [4] is not a new technology, but it has not gained public attention until recently. The ultimate

goal of designing a handwriting recognition system with an accuracy rate of 100% is quite illusionary, because even human beings are not able to recognize every handwritten text

without any doubt. For example, most people can not even read their own notes. Therefore there is an obligation for a writer to write clearly. In this paper, both Pattern Recognition and Neural Networks [2] will be defined. Examples of types

of Pattern Recognition and Neural Networks will be discussed. The advantages of using Neural Networks][2]tor ecognize handwritten characters will be listed. Finally, Artificial Neural Networks, using back-Propagation method will be used to train and identify handwritten digits.

2. PATTERN RECOGNITION

2.1. What is Pattern Recognition?

Pattern recognition system consists of two-stage process. The first stage is feature extraction and the second stage is classification. Feature extraction is the measurement on a population of entities that will be classified. This assists the classification stage by looking for features that allows fairly easy to distinguish between the different classes. Several different features have to be used for classification. The set of features that are used makes up a feature vector, which represents each member of the population. Then, Pattern recognition system classifies each member of the population on the basis of information contained in the feature vector.The following is an example of feature vectors that have been plotted on a graph.

Figure 2.1 Two classes fully separated

From Figure2.1, it shows two clusters of points. Each of them corresponds to one of the classes. These classes are fully separated and can be easily distinguished.

2.2. Pattern recognition methods.

Bayesian decision theory.

The Bayesian decision theory is a system that minimizes the classification error. This theory plays a role of a prioi. This is when there is priority information about something that we would like to classify. For example, suppose we do not know much about the fruits in the conveyer belt. The only information we know is that 80% of the fruit in the conveyer Malothu Nagu et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1685-1692 1685.

belt are apples, and the rest of them are oranges. If this is the only information we have, then we can classify that a random fruit from the Conveyer belt is apple. In this case, the prior information is the probability of either an apple or an orange is in the conveyer belt. If we only have so little information, then we would have the following rule: Decide”apple‘if P (apple) > P (orange), otherwise decide”orange‘ Here, P (apple) is the probability of being an apple in the conveyer belt. This means that P(apple) = 0.8 (80%). This is probably strange, because if the above rule is used, then we are classifying a random fruit as an apple. But if we use this rule, we will be right 80% of the time. This is a simple example and can be used to understand the basic idea of pattern recognition. In real life, there will be a lot more information given about things that we are trying to lassify. For example, we know that the color of the apples is red. Therefore if we can observe a red fruit, we should be able to classify it as an apple. We can have the probability distribution for the color of apples and oranges. Let wapp represent the state of nature where the fruit is anapple, let wora represent the state of nature where the fruit is an orange and let x be a continuous random variable that represents the color of a fruit. Then we can have the expression p(x|wapp) epresenting the density function for x given that the state of nature is an apple. In a typical problem, we would be able to calculate the conditional densities p (x|wj) for j so it will be either an apple or an orange. We would also know the prior probabilities P(wapp) and P(wora). These represent the total number of apples versus oranges in the conveyer belt. Here we are looking for a formula that will tell us about the probability of a fruit being an apple or an orange just by observing a certain color x. If we have the probability, then for the given color that we observed, we can classify the fruit by comparing it to the probability that an orange had such a color versus the probability that an apple had such a color. If we were more certain that an apple had such a color, then the fruit would be classified as an apple. So, we can use Baye's formula, which states the following:

$$P(wj |x) = p(x|wj) P(wj)/p(x)$$

What the formula means is that using a priori information, we can calculate the a probability of the state of nature being in state w hat we have given that the feature value x has been measured. So, if we observe a certain x for a random fruit in the conveyer belt, then by calculating P(wapp/x) and P(worg /x). we would decide that the fruit is apple if the first value is greater than the second one and if P(worg/x) is org greater, then we would decide that the fruit is orange. So, the Bayesian decision rule can be stated as: Decide worg if P(worg |x) > P(wapp |x), otherwise, decide wapp ,Since p(x) occurs on both sides of the comparison, the rule can also be equivalent to the following rule: Decide worg if p(x|worg)P(worg) > p(x|wapp)P(wapp), otherwise decide wapp The following graph shows the a posteriori probabilities for the two-class decision problem. For every x, the posteriors has to sum to 1. The red region on the x axes represents the values for x for which would decide as ”apple‘. The orange region represents values for x for which would decide as”orange‘.

2.3.A posteriori probabilities for two class decision

problem.

The probability that we would probably make an error is any minimum of the 2 curves at any oint, because it represents the smaller probability that we did not pick. The formula for the error is the following:

P(error|x) = min[p(wapp |x), p(worg |x)].

Nearest Neighbor rule.

The Nearest Neighbor (NN) [10]rule is used to classify andwritten characters. The distance measured between the two character images is needed in order to use this rule. Without a priori assumptions about the distributions from which the training examples are drawn, the NN [2]rule achieves very high performance. The rule involves a training set of both positive and negative cases. A new sample is classified by calculating the distance to the nearest training case. The sign of the point determines the classification of the sample. The following figure shows an example of NN

The Neural Network rule.

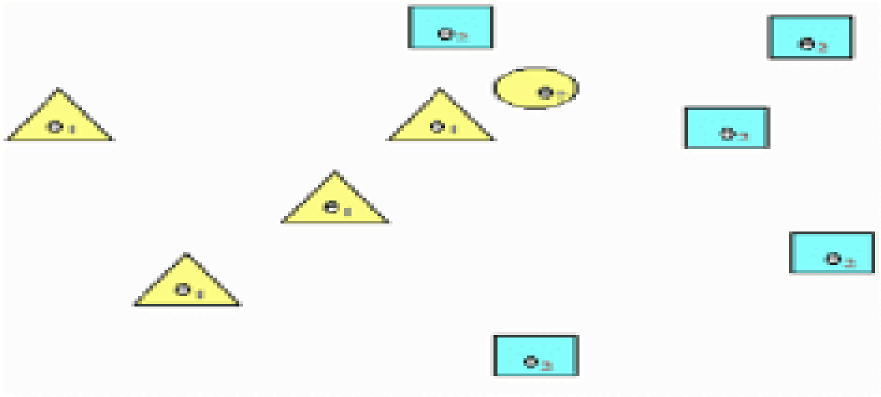
In this example there are two classes θ1, which are the yellow triangles and θ2 which are blue squares. The yellow circle represents the unknown sample X.We can see that the unknown samples nearest neighbor is from class θ1 therefore it is labeled as class θ1.When the amount of pre- lassified points is large, it is good to use the majority vote of the nearest k neighbors instead of the single nearest neighbor. This method is called the k nearest neighbor (k-NN) rule. The k-NN[2] rule extends the idea by taking and assigning the k nearest points the sign of the majority. It is ommon to choose k small (with respect to the number of samples) so thatthe points are closer to x to give accurate estimate of the true Malothu

Nagu et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1685-1692 1686 class of x.If the k values are large, it will help reduce the effects of noisy points within training data set. Also, large k values will minimize the probability of misclassifying x. The following figure shows an example of k-NN rule with k value equal 3:

Figure 2.4 The Nearest k-Neural Network rule with k=3.

As before, there are two classes: θ1 which are the yellow triangles, and θ2 which are the blue squares. The blue circle represents the unknown sample x. We see that two of its nearest neighbors are from class θ2 , so it is labeled as class θ2.

le



igure 2.3 The Neural Network rule

2.2.3 Linear Classification Discrimination.

The goal of Linear Classification is to assign observations into the classes. This can be used toestablish a classifier rule so that it can assign a new observation into a class. In another words, the rule deals with assigning a new point in a vector space to a class separated by a boundary. Linear classification provides a mathematical formula to predict a binary result. This result is a true or false (positive or negative) result or it can be any other pair of characters. In general, we will assume that our Results are Boolean variable. To do this prediction, we use a linear formula over the given input data. We refer this as inputs. The linear form is computed over the inputs and the result is compared against a basis constant. It is depending on the result of the comparison that we would be able to predict true or false. The following is the equation that can be stated as the discriminator: a1 x1 + a2 x2 + ... + an xn > x0 Here, a1, a2 are the variables that correspond to one observation. and x1,x2 together with x0 are the solution vector plus the basis constant. orresponding to each of the input vector a, there is a variable b. This variable b is the dependent Boolean variable. We refer this as the output. We normally have many m data points (a row vector of inputs a and the corresponding output, b). For convenience, we place all the data in matrices andvectors. A denotes the matrix of the inputs and it is in the dimension of m by n. The m Boolean outputs will be storedin a vector B. Now, the problem of linear classification can be stated in terms of the matrices and vectors. For example: we want to find x and x0 such that: (Ax > x0) equiv B This equivalence means that every row of the left-hand-side is a relation. The relation for the given data will be true or false. This has to match the corresponding B value. When we have more data points than columns, this is m > n, the problem does not have a solution. We can not find a vector x that will satisfy all the true values. Therefore we try to find a solution that can match vector B as good as possible. The best matching means that there is a minimum number of errors or there is a weighted minimum of errors. There will be a weight that is assigned to the true errors and nother weight that is assigned to the false errors. The idea of obtaining good classification is to be able to use it to predict the output for new data points. Here, the discriminator is a prediction formula, A and B are the training data, and x and x are the parameters of the model fitted to the training 0 data. The following figure shows the main components of linear classification:

2.5 Linear Classification in two dimensions.

Here, true are marked with a green T and false are marked with a red F. The inputs are two dependent variables, a1 and a2 . Here we use them to position the points in the plane. There are eleven Ts (positives) and twelve Fs (negatives).This linear classification in two dimensions is a straight line.The straight line drawn is a very good classification because the line separates most of the points correctly. We can see that all the Ts are on one side of the graph and all the Fs, but one, are on the other side of the graph. Therefore this classification has one "false positive" and no "false negatives". A false positive is a data point that is classified as positive but it is negative and vice versa for false negatives.

3. HANDWRITTEN CHARACTER RECOGNITION IN PATTERN RECOGNITION.

Linear Classification [9] is a useful method to recognize handwritten characters.[4] The ackground basis of Artificial Neural Network (ANN)[2] can be implemented as a classification function. Linear Classification works very similar to Artificial Neural Network because the apping of the ANN cell or one layer of the ANN cell is equivalent to the linear discrimination function. Therefore, if the ANN is a two-layer network, which is consisting of an input and an output layer, it can act as a linear classifier.

What is Neural Network?

A Neural Network (NN) [2]is a function with adjustable or tunable parameters. Let the input to a neural network bedenoted by x. This is a real-valued or row vector of lengthand is typically referred to as input or input vector or regressor or sometimes pattern vector. The length of the vector x is the number of inputs to the network. So let the network output be denoted b Y. This is an approximation of the desired output y, which is also a real-valued vector having one or more components and the number of outputs from the network. The data sets often contain many input and output Malothu Nagu et al, / (IJCSIT) International Journal of Computer Science and nformation Technologies, Vol. 2 (4) , 2011, 1685-1692 1687 pairs. The x and y denote matrices with one input and one output vector on each row.A neural network[2] is a structure involving weighted interconnections between neurons or units. They are often non-linear scalar transformations but can also be linear scalar transformation. The following figure shows an example of a one-hidden-layer neural network with three inputs, x = {x1,x2 ,x3}.

Figure 3.1 Feed-forward Neural Network with 3 inputs, two

hidden neurons and one output neuron. The three inputs, along with a unity bias input, are fed each of the two neurons into the hidden layer. The two outputs from this layer and from a unity bias are then fed into the singleoutput layer neuron. This produces the scalar output Y.Thelayer of neurons is called hidden layer because the outputs are not directly seen in the data.Each arrow in the Figure 3.1 corresponds to a real-valuedparameter, or a weight, of the network. The values of these parameters are tuned in the training network.A neuron is structured to process multiple inputs. This includes the unity bias in a non-linear way. Then, this produces a single output. All inputs to the neuron are first augmented by multiplicative weights. These weighted inputs are summed and then transformed via a non-linear activation function and as indicated from the above Figure 3.1, theneurons in the first layer of the network are non-linear. The single output neuron is linear because no activation function is used. The information in an ANN is always stored in a number of parameters. These parameters can be pre-set by the operator or trained by presenting the ANN with example.

region represents values for x for which would decide as "orange".

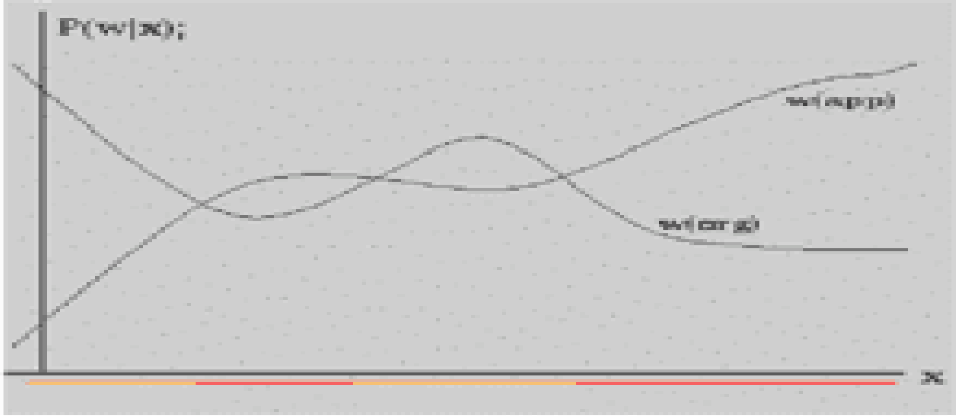


Figure 2.2 A posteriori probabilities for two class decision

3.2 Artificial Neural Network.

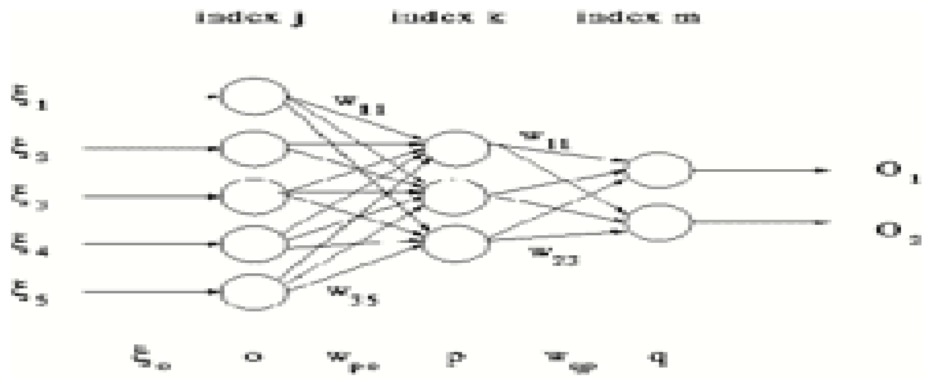
Artificial Neural Network (ANN) has been around since the late 1950's. But it Was not until the mid-1980 that they became sophisticated enough for applications. Today, ANN[2] is applied to a lot of real- world problems. These problems are considered complex problems. ANN‘s are alsoa good pattern recognition engines and robust classifiers. They have the ability to generalize by aking decisions aboutimprecise input data. They also offer solutions to a variety of classification problems such as speech, character and signal recognition.Artificial Neural Network (ANN) is a collection of verysimple and massively interconnected cells. The cells are arranged in a way that each cell derives its input from one or more other cells. Itis linked through weighted connections to one or more other cells. This way, input to the ANN is distributed throughout the network so that an output is in the form of one or more activated cells. es of input and also possibly together with thedesired output. The following figure3.2 is an example of a simple of ANN:

Multi-Layer Artificial Neural Networks.

3.3. Examples of types of Neural Networks. 3.3.1. Multi-Layer Feed-forward Neural Networks.Multi-Layer Feed-forward neural networks (FFNN)[9] have high performances in input and output function approximation. In a three-layer FFNN, the first layerconnects the input variables. This layer is called the input layer. The last layer connects the output variables. This layeris called output layer. Layers between the input and output layes are called hidden layers. In a system, there can be more than one hidden layer. The processing unit elements arecalled nodes. Each of these nodes is connected to the nodes of neighboring layers. The parameters associated with node connections are called weights. All connections are feed forward; therefore they allow information transfer from previous layer to the next consecutive layers only. For example, the node j receives encoming signals from node i in the previous layer. Each incoming signal is a weight. The effective incoming signal to node j is the weighted sum of all incoming signals. The following figures (Figure 3.3 (a)) are an example of a usual FFNN and nodes (Figure 3.3 (b)):

Figure 3.3 (a) Three layers feed-forward neural net,

(b) Processing unit element. .3.2 Back-propagation algorithm. Back-propagation algorithm[11] consists of two phases. Firstphase is the forward phase. This is the phase where the activations propagate from the input layer to the output layer. The second phase is the backward phase. This is the phase where then the observed actual value and the requested nominal value in the output layer are propagated backwards so it can modify the weights and bias values. The following figure 3.4 is an example of the forward propagation and backward propagation. alothu Nagu et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1685-1692 1688



re 3.1 Feed-forward Neural Network with 3 inputs