# Letter Recognition Data Using Neural Network

**Hussein Salim Qasim**

**Abstract**

The letters dataset from the UCI repository website form a relatively complex problem to classify distorted raster images of English alphabets. In contrast to rather complex networks, difference boosting algorithm, a computationally less intensive Bayesian classifier, is found to produce comparable or better classification efficiency on this problem. The character images were, originally, based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. We typically chose, randomly, 1,000 unique stimuli for study. We made sure that the distribution remains the same after choosing the one thousand stimuli. In this study, a neural network tool was developed for the purpose of predicting to identify each of a large number of black and white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 1,000.

**Keywords:** letter Recognition, Neural Network, Prediction, UCI Repository, Holland-style, Statlog-Project, English Alphabet, Backpropagation.

## 1.0 Introduction

Letter Image Recognition Data is created by David J. Slate since January, 1991. Originally, he created these data to investigate the ability of several variations of Holland-style adaptive classifier systems to learn and correctly guess the letter categories associated with vectors of 16 integer attributes extracted from raster scan images of the letters. Research results and experience worldwide provides clear evidence is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet, this is very important for every one use English language.

**Past usage:**

The database was originally created for research that investigated the ability of several variations of Holland-style adaptive classifier systems as mentioned before. The best accuracy obtained was a little over 80%. The first 16000 examples were used to train, while the remaining 4000 where used to test.

Another project using this database was the Statlog-project.

The parent font represented a full range of character types including script, italic, serif and gothic. The features of each of the 1,000 characters were summarized in terms of 16 primitive numerical attributes. The attributes are:

x-box: horizontal position of box,
y-box: vertical position of box,
width: width of box,
high: height of box,
onpix: total # on pixels,
x-bar: mean x of on pixels in box,
y-bar: mean y of on pixels in box,
x2bar: mean x variance,
y2bar: mean y variance,
xybar: mean x y correlation,
x2ybr: mean of x * x * y,
xy2br: mean of x * y * y,
x-ege: mean edge count left to right,
xegvy: correlation of x-ege with y,
y-ege: mean edge count bottom to top,
yegvx: correlation of y-ege with x].

For example, the horizontal position, counting pixels from the left edge of the image, of the center of the smallest rectangular box that can be drawn with all on pixels inside the box while the vertical position, counting pixels from the bottom, of the box. All attributes values are integers (1-15). The letters interval is (1-26 values) which means (A to Z).

Hence, in this study, a neural network tool was developed for the purpose of predicting to identify each of a large number of black and white rectangular pixel displays as one of the 26 capital letters in the English alphabet.

Research examined the effects of different procedures for encoding attributes, deriving new rules, and apportioning credit among the rules. Binary and Gray-code attribute encodings that required exact matches for rule activation were compared with integer representations that employed fuzzy matching for rule activation. Random and genetic methods for rule creation were compared with instance-based generalization.

The inputs for the neural networks are a modified subset of the numerical attributes which have been used by D.J. Slate in the

Letter Image Recognition data collection and we used 16 attributes (Frey, et al. 1991)

### 1.1 Neural Networks

DARPA Neural Network Study (1988) define a Neural Network (NN) as: "a neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes.", NN technique are consists from three layers, input layer, hidden layer and output layer, see Figure 1.

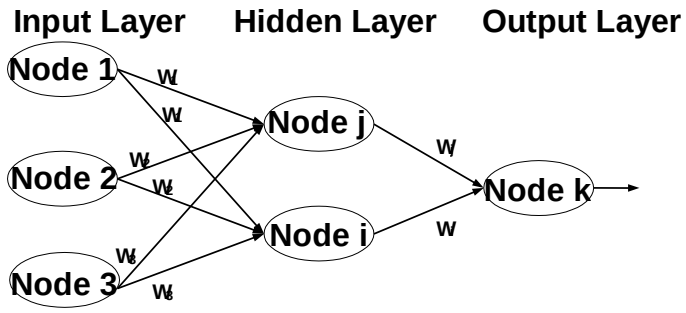


**Figure1:** Example of Neural Network architecture

The input and output layers must be numeric prefer to consider constraining to (0-1) range. Thus if we have Categorical input format, we must convert it to numerical values and scale it to (0-1) range. Furthermore the Weight is performance parameters of the feed-forward neural network in hidden layer. The training algorithm of the ANN (Artificial Neural Net) is affected by, starting with arbitrary weights, presenting the data, instance by instance, adapting the weights according the error for each instance, and Repeating until convergence (Brookshear, 2005). The backpropagation algorithm updates the weights using the difference of actual response and the function response for each instance.

Neural Networks (NNs) are networks of neurons, for example, as found in real (i.e. biological) brains (bullinarie, 2004)

In this study, a neural network tool was developed for the purpose of predicting to identify each of a large number of black and white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The system integrates Neural Network with Backpropagation learning algorithm to establish a prediction model to be applied in predicting Letter Recognition choice that uses Neural Network with Feedforward algorithm which is performed through online

## 2.0 Letter Recognition Data system

The Letter Recognition Data system involves neural network approach to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet, from correctly guess the letter categories associated with vectors of 16 integer attributes extracted from raster scan images of the letters.

The system integrates various NN technologies, techniques and data analysis using Visual Basic environment using minimal hardware and software requirements.

### 2.1 Data Preparation

To construct a prediction model, a data set containing 1000 of records were trained. The data was taken from Odesta Corporation; 1890 Maple Ave; Suite 115; Evanston, IL 60201 David J. Slate (January, 1991) The samples character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15.

| Attributes | Data Type | Range | | |
|---|---|---|---|---|
| | | Min | Max | Set |
| letter | Character | | | A To Z |
| x-box | Integer | 0 | 15 | |
| y-box | Integer | 0 | 15 | |
| width | Integer | 0 | 15 | |
| high | Integer | 0 | 15 | |
| onpix | Integer | 0 | 15 | |
| x-bar | Integer | 0 | 15 | |
| y-bar | Integer | 0 | 15 | |
| x2bar | Integer | 0 | 15 | |
| y2bar | Integer | 0 | 15 | |
| xybar | Integer | 0 | 15 | |
| x2ybar | Integer | 0 | 15 | |
| xy2bar | Integer | 0 | 15 | |
| x-ege | Integer | 0 | 15 | |
| xegvy | Integer | 0 | 15 | |
| y-ege | Integer | 0 | 15 | |
| yegvx | Integer | 0 | 15 | |

**Table 1.0:** Attributes and data type

## 2.2 System Development

Letter Recognition Data system is developed using several programming languages and development tools such as Visual Basic, Microsoft Excel. Letter Recognition Data system comprises of two main modules: training module and prediction module. Fig. 2 depicts the interactions of both modules.

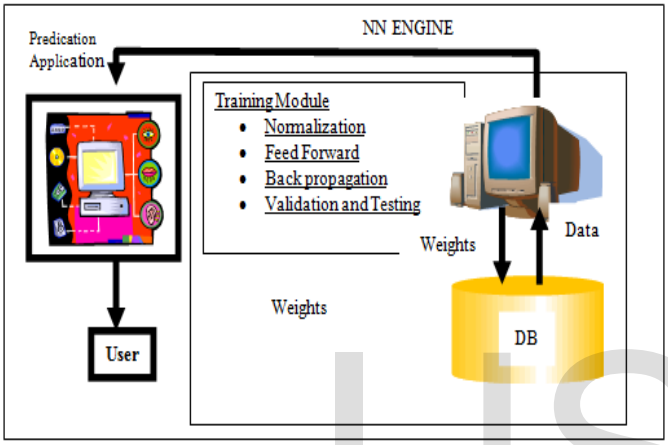


**Figure2:** Letter Recognition Data Using Neural Network

## 2.3 Training module:

This module deals with learning from a number of inputs to classify output. The learning is performed using one of Neural Network algorithms known as Backpropagation. It produces a model with a set of weights. These set of weights explain the association values between variables in classifying a set of inputs into an output. This module allows parameter setting and also data preprocessing, to include cleansing prior to training, testing and validating (see Fig. 3.0 and Fig. 4.0).

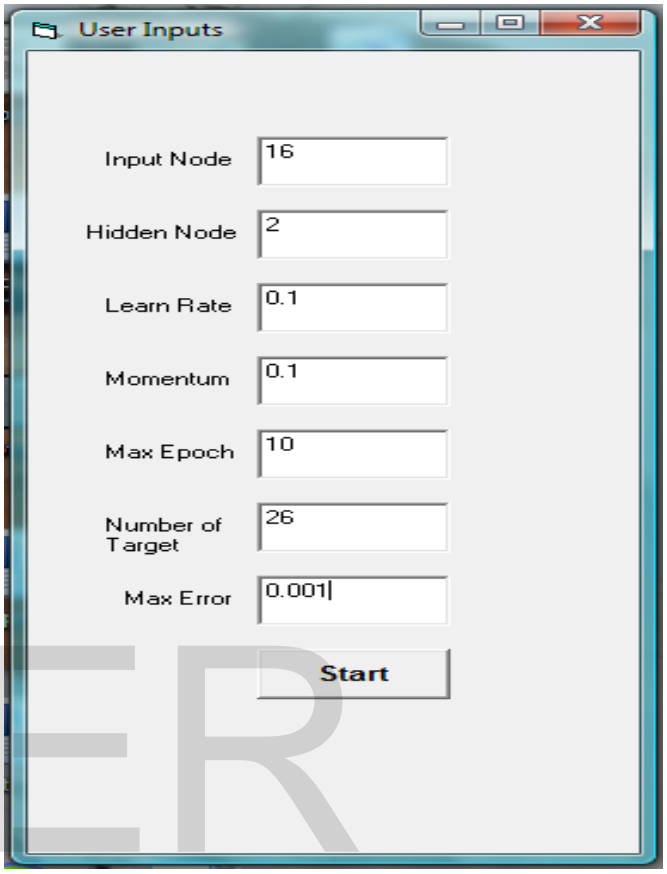


**Figure 3.0:** Parameter setting interface

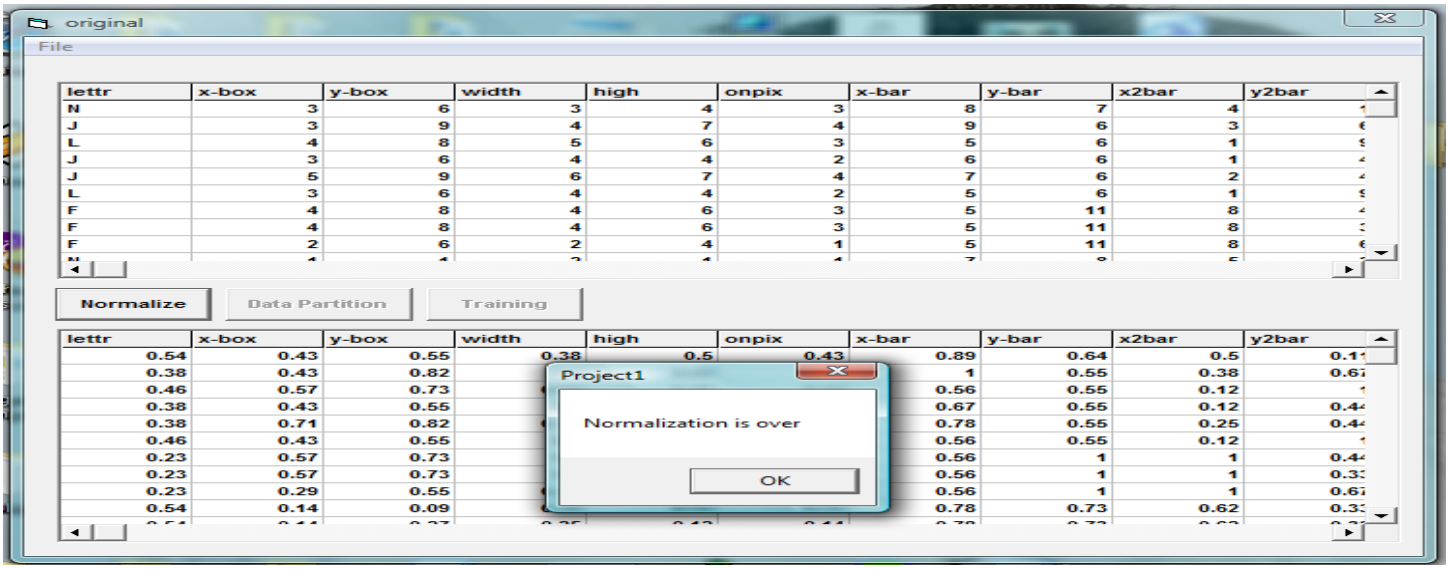


**Figure 4.0:** Data preprocessing interface

NN model is obtained by varying the number of hidden units, the learning rate, the momentum and the stopping criteria. For these purposes, the data set is partitioned into three groups; training (60%), testing (20%) and validation (20%). Fig. 5.0 describes the training interface.

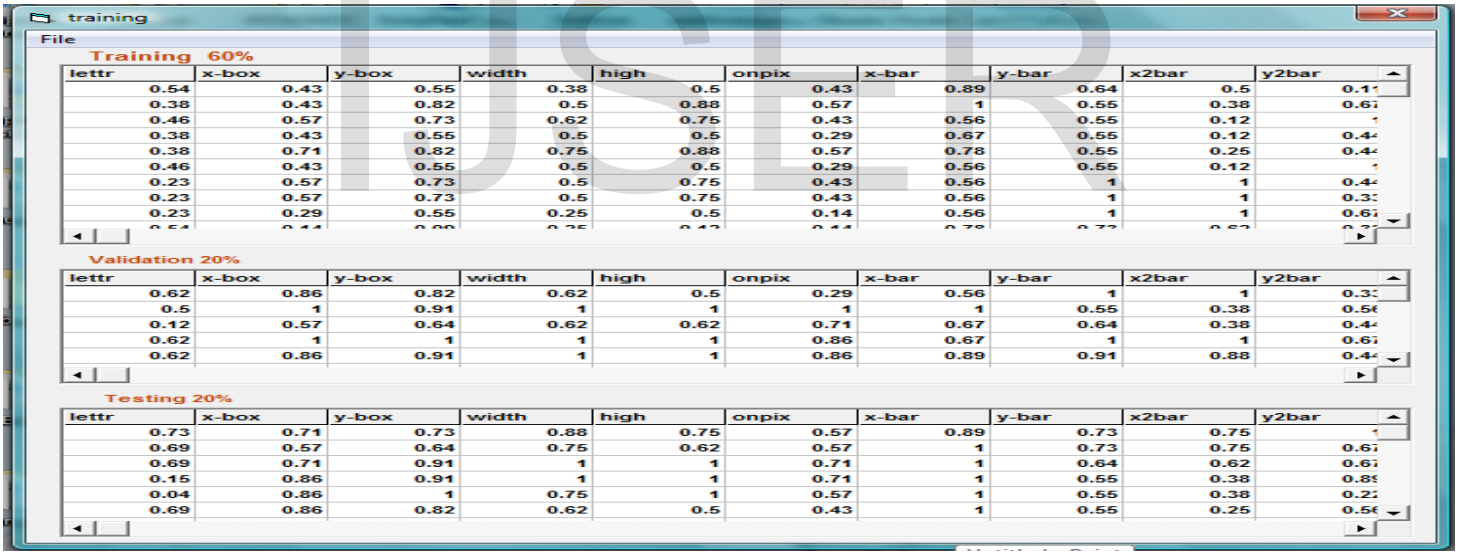


**Figure 5.0:** Training interface

As the training finished after certain period or condition, the system gives output that state the percentages of training, validation and testing correctness together with the extracted weights from the learning process (see Fig. 6.0).

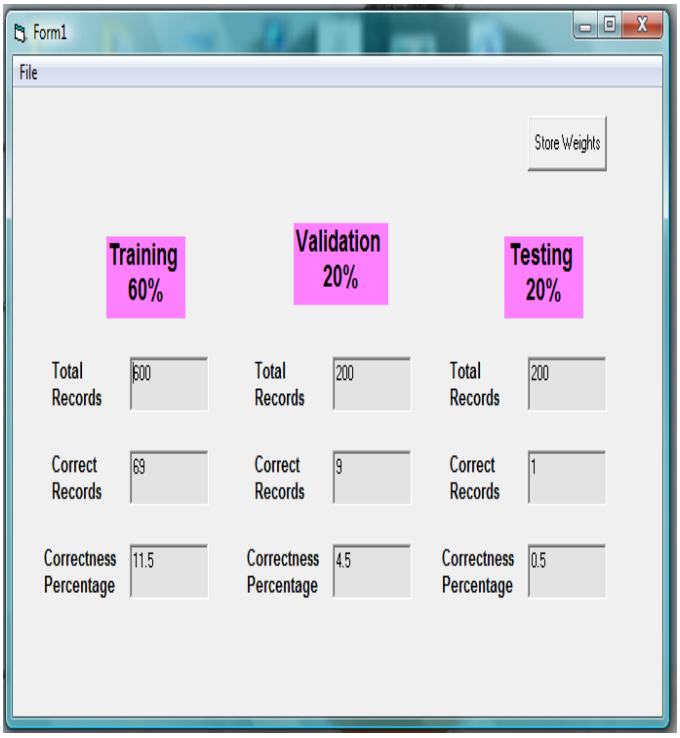


**Figure 6.0:** Training output interface

2.4 **Prediction module:**

Once the prediction model is obtained from the training module, Letter Recognition Data Using Neural Network system is able to recognize one of the 26 capital letters in the English alphabet by identifying each of a large number of black and white rectangular pixel and using a feedforward Neural Network algorithm. Before doing prediction, the user must fill in all the attributes within the given range. The insert window and result value is shown below (fig. 7).

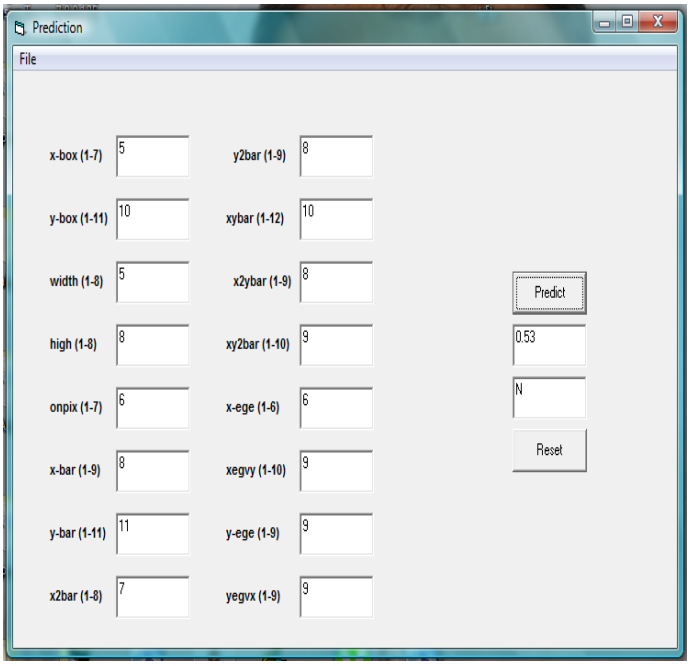


**Figure 7.0:** Prediction result

**3.0 Results**

The Letter Recognition Data Using Neural Network system used a data set containing sixteen of integer attributes extracted from raster scan images of the letters preprocessed, trained, tested and validated. A total of 1000 samples of data were preprocessed before the data is fed into training module for training, testing and validating. The testing result indicates that NN obtained 60 % prediction accuracy. Table 2.0 details the predictive model established by the training module.

| Parameter | Value |
|---|---|
| Neural Network Architecture | Multilayer perceptron |
| Learning Algorithm | Backpropagation |
| Num. of inputs | 16 |
| Num. of hidden units | 2 |
| Num. of output classes | 26 |
| Learning rate | 0.1 |
| Momentum rate | 0.1 |
| Stopping criteria | 10 epochs |

**Table 2.0:** Neural Network model for Predicting the suitable contraceptive method choice

The prediction module is designed in visual basic, thus the system can be used by anyone who interested to get some information regarding their letter recognition specifically to identify each of a large number of black and white rectangular pixel displays as one of the 26 capital letters in the English alphabet.

**4.0 Conclusion and Recommendation**

In this study, a neural network tool was developed for the purpose of predicting the letters. Hence, the result shows that developed system can be used as a tool to assist identifying each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. All 26 classes have almost the same number of samples.

For any single feature, the in-class variance is almost same as the between-class variance. Hence it is difficult to distinguish 26 classes by only few features. We used a subset for training and another subset for validation, the third one for testing. We got the result for training is 11.5, the validation is 9.5 and the result for testing is 0.5.

The findings in this study imply that NN has potential to be used to find a prediction tool in letter recognition, and we can use it in tutorial purpose in school or college.

Letter Recognition Data can help the public to know kinds of letters. In addition, the system also suggests the suitable letter for a person to know any letter he/she want to use, the system can be improved by using the whole 20,000 unique stimuli based on 20 different fonts.



**References**

Brookshear, J.G., (2005). Computer Science An Overview, Chapter 10, Eighth Edition.

bullinaria, J. A., (2004). Introduction to Neural Networks,Birmingham, B15 2TT, UK ,. Pp.L1-9

David, J., Slate, (1991). Letter Image Recognition Data, Machine Learning Vol 6 #2.

Frey, P. W., Slate, D. J., (1991). Letter Recognition Using Holland-style Adaptive Classifier, in Machine Learning Vol. 6, No. 2, pp. 161-182.

Kaynak, C., (1995). Methods of Combining Multiple Classifiers and TheirApplications to Handwritten Digit Recognition, MSc Thesis, Institute of Graduate Studies in Science and Engineering, Bogazici University.

Leenheer,P. D., Aabi,M., (2001-2002). upport Vector Machines: Analysis of its Behavior & xtension for Large-Scale Problems, pp. 128

- Hussein Salim Qasim , Msc. Information Technology / University Utara Malaysia, work as Lecture Assistant in College of Pharmacy / Almustansiriyah University. Husain_salem@yahoo.com