Threshold Value of Activation Function Changes the Effectiveness of Neural Network and deep learning in the Research of Letter Recognition

Zhijing Ke

Research School of Computer Science,

Australian National University

u6343103@anu.edu.au

Keywords: neural network, machine learning, ReLU, Threshold Value, letter recognition, Oversampling, classification, deep neural net,

Abstract

Activation function for hidden layers differs from neural networks in effectiveness, speed and applicability, and threshold value of activation function can improve the algorithm's result by improving activation function's effect. The effect exists but are usually difficult to generalized for the condition of the algorithm training is difficult to meet the ideal research condition. Deep learning was inspired by biological nervous system, aiming to process layering and abstraction tasks. Letter recognition data is about recognizing capital letters in different fonts in same specification images[1], which provides an idea to recognize letters, numbers and symbols. The letter recognition dataset this report used had a high stability of accuracy while the train data was select randomly, which made it desirable to generalize the effect of threshold value. Some methods and pre-process, including normalization and raw data' transformation also had been used in this research. At the end the result of the research will be compared with relevant research result in the algorithm accuracy[2], and the research target is to generate a primary conclusion of how the threshold value influences the effectiveness of single layer neural network and deep neural network. The result was that threshold value can improve the result of the algorithm in some cases, while the effect in deep learning was unsatisfactory.

Introduction

The reason for choosing letter recognition is this model is a classification task and has a high stability accuracy in the result, which provide a desirable condition to search the importance of threshold value. Deep neural network can also be discussed for the features and abstractions for the letters is interesting. This model also provided a unique idea to process simple image data like letters, symbols and numbers, which might be useful in scanning and recognizing hand writing information in the future. The data attributes can be normalized well and enable to be used in neural network training directly. All the attributes are effective and ideal. The 16 attributes are normalized to the value between 0 to 15 and they are the details of image data of 20 different font distorted letters being transformed to black-and-white images, such as the height of the image or the weight of the image, and the position of black area.

This research will use neural network and deep learning to try to predict the specific letter for each image data and use several methods to improve the prediction function with some changes of activation function for neural network, oversampling and under sampling, try to find an elementary conclusion of how the change the threshold values influence the effectiveness of the neural network. Pre-process are also used while processing raw data.

Method

Pre-process

The dataset of this research needs to be pre-processed considering the target prediction values are capital letters between "A" to "Z". In the algorithm this column will been transformed in to ASCII code and then generated into integer "0" to "25". This will let the output meet neural network's output requirement. Then algorithm will change the data's type into numeric. Length of the dataset after pre-process is 20'000 and this will be divided randomly into a 16'000 size train-dataset and a 4'000 size test-dataset.

Sampling is significant for training algorithm. Imbalanced data in real world are generally recorded in order and this will cause invalid learning for some machine learning and neural network algorithms for some of the minority of some values in target values did not trained well in the algorithm [3]. That will lead algorithm enable to predict some minor target values. This will be serious for in most cases those values are more possible to be what the research wants to recognize. Oversampling and under sampling is a reasonable way for sampling and train the algorithm can be trained effectively. By calculate each value between "0" to "25" we can generally know the result's distribution. For those minor values in the target we do oversampling and for those major values we do the under sampling. Then the train data set will become a different rate from the target values classes. Outliers will also influence the effectiveness of the algorithm. Due to the data are already normalized, the length of the dataset did not decrease after several outliers removing way had been used, such as bimodel distribution removal[4] and inter quartile range removal. After pre-process the length of data is still 20'000.

The result has 26 levels, so it will be difficult to create a confusion matrix, relevant index were also difficult to be concluded. Accuracy is more important in this algorithm. Besides, ROC curve also had been used in this research.

Threshold Change

The activation function in neural network is significant. For hidden layers, sigmoid function is a common function that can be used as activation function. Other activation function such as linear rectified linear unit (ReLU), which could be used for handling vanishing gradient problem, can also used in hidden layers. In this research, the algorithm will use ReLU and sigmoid function for hidden layer. When using ReLU activation function, the algorithm is faster than sigmoid function and Multiple hidden layers neural network. Multiple hidden layers also had been tried though the learning effectiveness did not rise, and the model even began to degeneracy. Changing the threshold for both activation function will influence the prediction accuracy.

For ReLU function, after different value of threshold had been tried, 0.05 might be a reasonable threshold value in the algorithm for this dataset.

$$X^h = X^f - 0.05 \qquad (1)$$

X^h means the forward hidden neurons' input and X^f means the output of the hidden neurons.

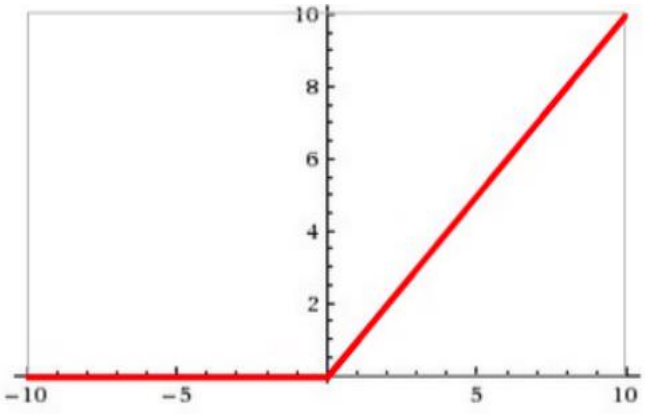

Figure 1. a visualization picture of ReLU function minus 0.05

```
h_input = self.hidden(x) -0.05
```

For Sigmoid activation function, the result will be discussed in the next section.

Deep Neural Network

After Changing the threshold value for simple artificial neural network, the research built a deep neural network with three hidden layers. ReLU function was used in the third hidden layer. Cross Entropy loss function can be used for the validity of the neural network. By training this deep neural network, the research can find interesting differences of accuracy between simple neural network with change of threshold value and deep neural network. The number of epoch was the same as the simple neural network above.

Deep Neural Network with Threshold Value

The next method was the combination of threshold value with deep neural network. The threshold value change was in the first hidden layer. The research also used ReLU as the activation function. The number of epoch was the same as the deep neural network above.

Result and Discussion

Besides the method the algorithm tried above, amount of parameters combinations had been tried. Finally, while hidden neurons equal 20, number of each epoch equals 5000 learning rate equals 0.02 for the simple neural network the prediction's accuracy of test dataset is reasonable enough. While the hidden neurons for deep neural network were 30, 22 and 15. The accuracy is lower while those three values are far from the range. Following are tables showed about the result.

Besides, the threshold also showed a rise of test data accuracy, due to the original parameters combination already had a high accuracy the rise of threshold is not very significant. While in a low accuracy, the improvement will be much better. In the different activation function, sigmoid function in this research is not very appropriate to change threshold: the accuracy decrease in all the situation with different parameters, and it is slower than the ReLU function.

The oversampling and under sampling also didn't raise the accuracy. That might be the original data itself is balanced and different ways sampling could not improve the algorithm and the accuracy. ROC curve and AUC showed the models were reasonable.

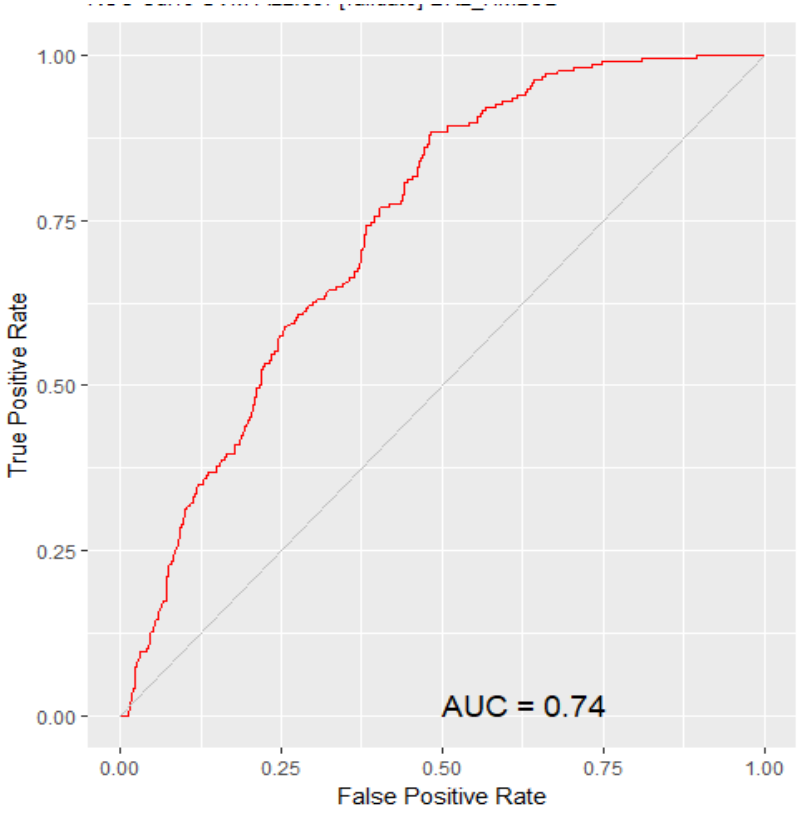

Figure 2. a visualization picture of ROC curve of DNN

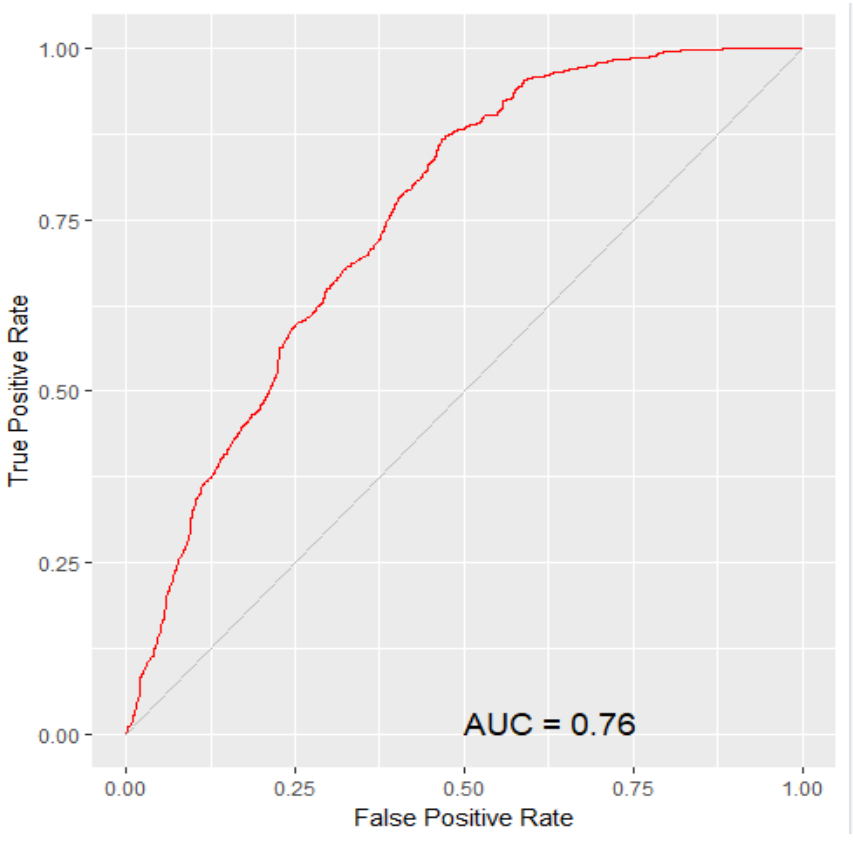

Figure 3. a visualization picture of ROC curve of simple neural network

Compared with Frey and Slate's result [1], our simple neural network's best accuracy is 8.1% higher, which was 82.7%. And the threshold change of simple neural network is 8.7% higher than their accuracy. The original deep neural network is 2.9% higher than their result, which was 87.64%. deep neural network's accuracy was 3.1% lower than their result.

Vanishing gradient problem might have happened in the deep neural network with threshold value. The same problem might also have happened in the original deep neural network, but maybe it was not obvious enough.

The deep neural network's performance was worse than the simple neural network, and deep neural network's performance were even worse. That might be the activation function's threshold value changed the learning validity of the neural net, some interesting patterns and methods were ignored by the activation function due to the change of threshold value.

During the research the threshold value did improve the algorithm though at the end it was not a majority improvement while the accuracy was reasonably high, and in other cases the result will not be satisfactory. This way was not appropriate to all activation function for hidden layer. Whether this way can improve the algorithm when the threshold value was not in the first hidden layer remains to be researched.

Table 1. Accuracy with change of epoch number and learning rate

Epoch_num	Learning rate	Accuracy
1000	0.002	66.31%
1000	0.1	69.58%
5000	0.002	90.81%
5000	0.01	70.46%
20000	0.002	88.92%
20000	0.01	77.61%

Table 2. Accuracy with change of threshold value

Threshold value	Accuracy
0.00	90.81%
0.01	90.76%
0.05	91.39%
0.1	87.23%

Table 3. Accuracy with different method of neural network

Method	Accuracy
NN	90.81%
NN with threshold	91.39%
DNN	87.64%
DNN with threshold	79.62%

Conclusion and Future Work

Overall, change of threshold value is useful in specific models and datasets. The threshold can improve the validity of the model to get more accurate result due to all the interesting methods concluded from the model are stronger than the original neural network. But it might be more possible that more interesting patterns and method had been ignored by the neural network, which will cause poor performance of the models. The same tech should not be used in deep learning in most cases, the abstraction of patterns will be influenced, and the result are much possible to be worse.

In the future, the algorithm should use other dataset and build neural network with multiple hidden layers and conclude the result of different threshold values in different hidden layers. Different activation function also should be tested to draw a more specific and more accurate conclusion. It is possible if some sampling methods can be used in this research.

References

1. Slade, P., and Tamás D.Gedeon: dimodal distribution removal. International Workshop on Artificial Neural Networks. Springer, Berlin, Heidelberg (1993)
2. Frey, P. W., & Slate, D. J: Letter recognition using Holland-style adaptive classifiers. Machine learning, 6(2), 161-182 (1991)
3. Mao, W., Jiang, M., Wang, J., & Li, Y.: Online Extreme Learning Machine with Hybrid Sampling Strategy for Sequential Imbalanced Data. Cognitive Computation, 9(6), 780-800 (2017)
4. Letter Recognition Data Set, http://archive.ics.uci.edu/ml/datasets/Letter+Recognition