Language User Prediction Using Neural Network with Bimodal Distribution Removal Algorithm and Genetic Algorithm

Xiao Jiang
U6683852 @anu.edu.au
Research School of Computer Science
Australian National University

Abstract. This paper builds a neural network based on the reading disturbance data set to predict whether the reader is a first language reader or a second language reader. The goal of this paper is to compare the optimization performance of Bimodal Distribution Removal algorithm and feature selection by using Genetic Algorithm. The BDR algorithm shows the remarkable results with the original results. The accuracy of the BDR algorithm is improved from 57.82% to 88.73%, and the optimization of the genetic algorithm is from 57.82% to 59.63%. It has found that the BDR algorithm is suited in this problem with this kind of neural network architecture, but the GA-based feature selection does not improve the performance of neural network because the size of this dataset is not enough.

Keywords: Bimodal Distribution Removal, Neural Network, Genetic Algorithm, Feature Selection

1. Introduction

Many studies have shown that it is effective to improve the training effect by transforming data, selecting features and optimizing algorithms. This research uses a search algorithm to solve the optimization: genetic algorithm (GA), which is essentially a random global search and optimization method developed based on the biological evolution mechanism of nature.[6] The bimodal distribution removal algorithm is also applied to this data set to develop and train neural networks.

The purpose of this study is to compare the differences between the genetic algorithm and the BDR algorithm. Firstly, preprocess the data and then apply the two methods to the randomly assigned training dataset and testing dataset. The performance of the two algorithms is evaluated by the accuracy and loss.

1.1 Background

Neural networks are widely used due to their satisfactory computing power. At present, the application of neural networks to the development of neural networks mainly uses the advantages of convolution operations and storage units. The bimodal distribution removal algorithm is a method that helps the training process by eliminating outliers, which may hinder the learning process.[2] But considering that some bad data will affect the learning of the neural network and these bad data will lead to a long-term fine-tuning process. Therefore, using BDR to delete these data or patterns can improve the entire training process. Genetic algorithm is a kind of evolutionary algorithm, and it is also a method to help reduce the training time of the model by selecting more relevant features to simplify the model.[1] Genetic algorithms perform feature selection through biologically inspired operations. Using selection, cross and mutate to find out features that are more adapted to the environment.[9]

2. Method

2.1 dataset description

This reading distraction dataset is a study on the problem of reading interference, which includes statistical data for first language readers and second language readers. It provides the relationship between reading behavior in two languages and the readers. L1 is English as the first language, and L2 is English as the second language. We train and predict whether the result is an L1 reader or an L2 reader by providing 20 influential features.

2.2 Data Processing

Through the analysis of reading distraction dataset, the *L1/L2* column is set as the target of each pattern. The values of *L1 / L2* are string values, so they are converted to binary numbers: local user is labeled as 0, and second language user is labeled as 1. The feature of complexity of reading are also converted to numbers, with simple text set to 0 and hard text set to 1. Finally, change other data (such as time) to decimal to facilitate calculation. After all these preparations, all the data will be normalized, because the different functions are originally in various ranges.

After preprocessing, there are 66 data records and 20 features. The dataset is randomly divided into training dataset and test dataset, 33%is used for test cases separately, and the 67% is used for training.

2.3 Network Architecture

The neural network is composed of input layer, hidden layer and output layer.[7] The input layer takes the preprocessed characteristics as input. The input data is a two-dimensional array with a shape of 66 * 20. Only one hidden layer was selected according to the size of this data set. The number of neurons in the output layer is 1, and the type of language reader is used as the output.

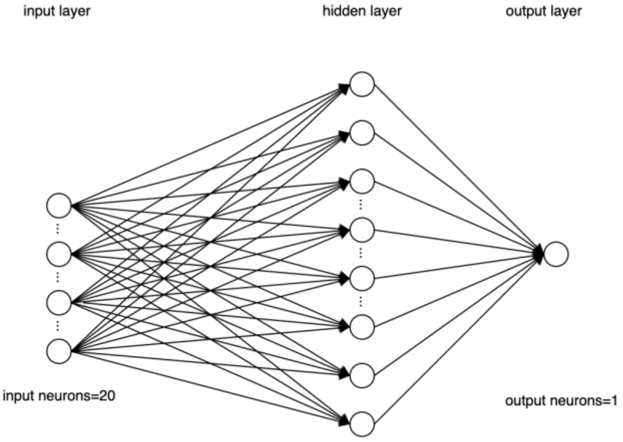

Figure 1: the neural network architecture

2.4 Evaluation Method

There are many methods for evaluating the performance of neural networks. This project uses the accuracy of training data sets and test data sets to evaluate the performance of neural networks. Another intuitive assessment is the convergence of the test data. Recording changes in losses can help identify the convergence of the training process. Ultimately, this report evaluates the performance of the neural network through accuracy and loss to evaluate the feature selection based on GA and BDR algorithm.

$$Accuracy = \frac{number\ of\ correct\ classified\ patterns}{total\ number\ of\ patterns} \tag{1}$$

2.5 Bimodal Distribution Removal

Bimodal distribution deletion (BDR) is used to delete noisy data points from valid data points. BDR can minimize the effects of severely errored patterns by removing noisy patterns from the data set.[2] It is an effective method that will improve the efficiency and accuracy of the neural network.

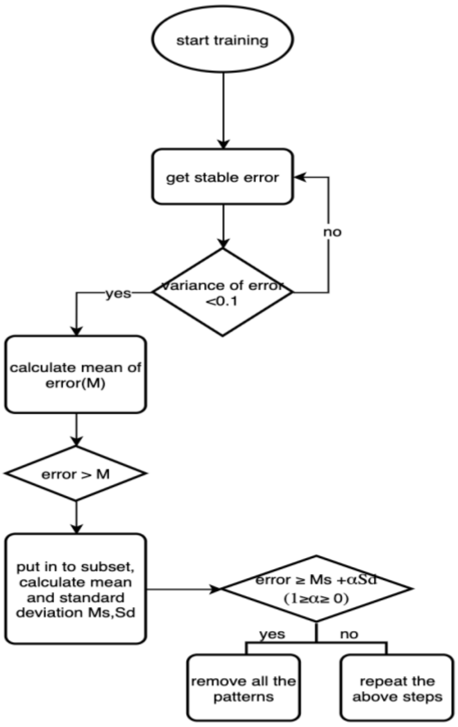

Figure 2: the flow chart of BDR algorithm

The specific step of the BDR algorithm is to train the entire training data set and start to calculate the average error M on the training set when the error variance is less than 0.1. If the error > M, then get these errors from the training data set and put them in the subset. Delete all patterns when error ≥ Ms + αSd (1≥α ≥0), the Ms is the mean and Sd is the standard deviation of the subset. Repeat these steps until the dataset error has dropped to the lower limit[2].

2.6 Feature Selection of Genetic Algorithm

Genetic Algorithm(GA) is an adaptive probabilistic search algorithm based on natural selection and genetic mutation.[8] In this algorithm, the individual is a binary string encoding and the individual is the main evolution object. Its simulated biological evolution has reproduction, crossover and mutation.[8]Breeding operations select the most suitable individual offspring from the population, while crossovers and mutations increase the diversity of the population. From an evolutionary perspective, the average adaptability of the new generation group to the environment is higher than that of the parental generation.

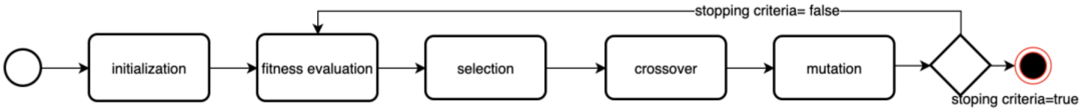

Figure 3: the flow chart of GA-based feature selection

2.6.1 Initialization

The first step in GA is to create and set the size of the initial population. In stochastic optimization methods, individual genes are usually initialized randomly. In this data set, each reader surveyed is regarded as a chromosome, and 20 features are the genes above. Genes are represented in a binary manner: 1 means that the feature is selected, and 0 means that the feature is removed. For example, as shown in the figure below, we assume that the first five features are x1, x3, x5 selected, then the chromosome is expressed as [1, 0, 1, 0, 1].

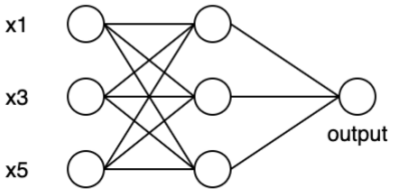

Figure 4: example of an individual with first 5 features

2.6.2 Fitness Evaluation

After the initialization, it needs to assign a fitness value to each individual in the population. The choice of fitness function directly affects the convergence speed of the genetic algorithm and whether it can find the optimal solution. Therefore, in this paper, neural network trained with the training instances and evaluated the error with the selection instances. A high selection error means a low fitness and these individuals with great fitness are more likely selected for recombination.[1]

2.6.3 Selection, crossover, mutation

Finally, it needs to go through the selection, crossover and mutation process. The selection operator is to select those individuals who are more adaptable to the environment, and the number is half of the total population. Usually this selection method is called Stochastic Universal Selection method.[9] Additionally, the crossover operator regroups the individuals selected in the previous steps to generate a new population. It will randomly select two individuals and combine their characteristics to generate offspring. As for mutation, it solves the problem of low diversity by randomly changing the values of certain features in the offspring. It needs to generate a random number between 0 and 1. When some value is less than this randomly generated mutation rate, these characteristics need to be mutated.[9]

3. Results and Discussion

3.1 Bimodal Distribution Removal Algorithm

3.1.1 Error Distribution

The use of the BDR algorithm is to remove noise pattern in the training dataset. As shown in the figure 5 below, the comparison between the original error distribution and the error distribution using BDR shows that the BDR algorithm can indeed remove the noise pattern.

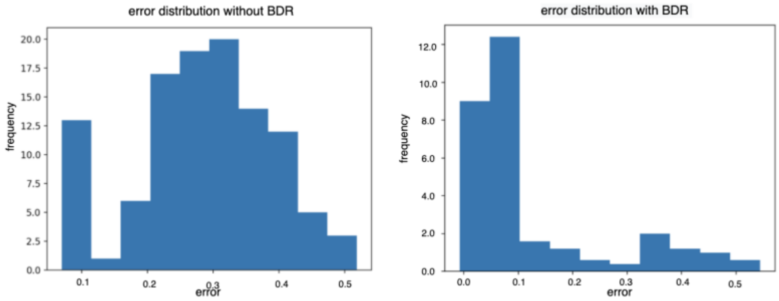

Figure 5: The Error Distribution comparing

3.1.2 Accuracy and Loss

The table and figures show the accuracy and loss of the training dataset and testing dataset with BDR. It is clear that they are differently with the initial results. The accuracy of training and testing improved much higher than that without BDR. As we can see, the accuracy of training improved from 51.28% to 89.87% with 500 epochs, and the accuracy of testing improved to 88.73% from 57.82%.

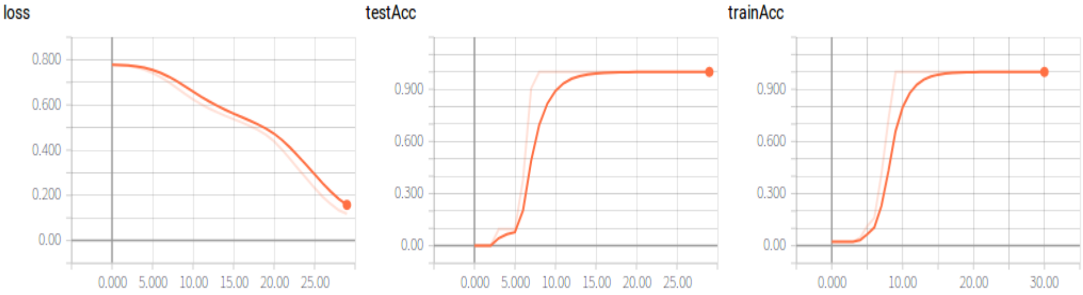

Figure 6: Training with BDR

Performance Parameter	Accuracy	Loss
Training accuracy%(500 epoch)	89.87	0.1675
Testing accuracy%(500 epoch)	88.73	0.1398
Training accuracy%(1000 epoch)	90.00	0.1248
Testing accuracy%(1000 epoch)	86.74	0.1182

Table 7: results with BDR

3.2 Feature Selection by GA algorithm

Feature selection is a method based on genetic algorithm to find the most suitable feature value. In the optimization process, population size, crossover probability and mutation probability are important parameters to be controlled. The final parameter values used were a population size of 100, a crossover probability of 0.75 and a mutation probability of 0.01. The adaptation value obtained as shown in the figure 8 below is between 60.04 and 60.05 and the value of fitness is not very high as the size of dataset is not big enough.

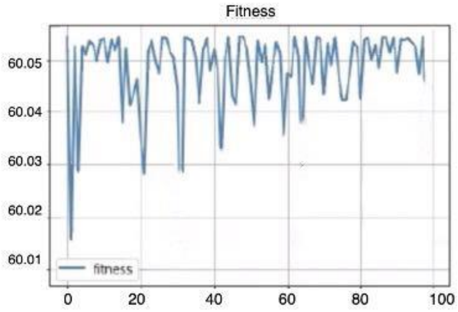

Figure 8: Fitness in GA

3.3 Comparison of BDR and GA

After comparing the results of features selected by the genetic algorithm with the BDR algorithm, it is found that after the genetic algorithm is used to optimize the features, the testing accuracy of the classifier is reduced. This is because the data set is too small to reflect the advantages of genetic algorithms. The results show that the BDR algorithm is more suitable for this dataset, which can effectively remove noise patterns to improve accuracy and reduce loss.

Method	Testing Accuracy(%)	Loss
Original result	57.82	0.2487
GA	59.63	0.2098
BDR	88.73	0.1398
BDR+GA	87.27	0.2282

Table 9: Result Comparison of Accuracy and Loss

4. Conclusion and Future Work

This report compares the performance of the BDR algorithm and the genetic algorithm based on the small sample dataset. From the results, it is concluded that the BDR algorithm has better performance than GA on this dataset. Firstly, the data is preprocessed and the accuracy of the initial results is 57.82%. After performing GA-based feature selection algorithm, the accuracy is improved to 59.63%. But BDR has good performance because it increases the accuracy to 88.73%. Finally, both the two algorithms are applied to the neural network. Although feature selection improves the training efficiency, the accuracy and loss are slightly lower compared with that using only BDR algorithm, which are 87.27% and 0.2282. Therefore, the conclusion is that the BDR algorithm is the most suitable for the current research of the dataset.

For further work, a more complex and huger dataset is needed here to study the performance of BDR algorithm and GA algorithm and then optimize them.

5. References

1. Fang, X. (2007). Engineering design using genetic algorithms.
2. Slade, P., & Gedeon, T. D. (1993, June). Bimodal distribution removal. In International Workshop on Artificial Neural Networks (pp. 249-254). Springer, Berlin, Heidelberg.
3. Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.
4. Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv: Learning,.
5. Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
6. Koul, S., & Chhikara, R. (2015). A hybrid genetic algorithm to improve feature selection. Int. J. Eng. Tech. Res, 4(5).
7. Mao, K. Z., Tan, K. C., & Ser, W. (2000). Probabilistic neural-network structure determination for pattern classification. IEEE Transactions on neural networks, 11(4), 1009-1016.
8. ElAlami, M. E. (2009). A filter model for feature subset selection based on genetic algorithm. *Knowledge-Based Systems*, *22*(5), 356-362.
9. Wu, Y. L., Tang, C. Y., Hor, M. K., & Wu, P. F. (2011). Feature selection using genetic algorithm and cluster validation. Expert Systems with Applications, 38(3), 2727-2732.