Comparison of BDR and GA on Backpropagation Network

Yihua Zhang

Research School of Computer Science
Australian National University
Canberra Australia
u6575450@anu.edu.au

Abstract. Many methods have been used in neural network to delete noisy. This paper is mainly to discover the performance of one of methods, Bimodal Distribution Removal (DBR)[1], applied for Backpropagation Neural Network (BP) on a dataset about mark prediction, and analyses its some advantages and disadvantages. By comparing to the results which do not involve BDR, BDR shows remarkable results, especially on the training dataset. The accuracy of training set is significantly improved from around 40% to at least 70%, while testing accuracy is not directly affected. As an extension, Genetic algorithm (GA) is applied for feature selection before training process. As a result, it does little help to improve the accuracy of training and testing.

Keywords: BDR, GA, Marks Prediction, BP

1 Introduction

Parameters of neural network model are always negatively affected by noise data in training set [1]. To improve the performance of model on testing data set, some methods must be found. As an effective approach, Bimodal Distribution Removal [2] proposed by Slade and Gedeon in 1993 focuses on permanently removing these input patterns that are filtered from training set by some calculated thresholds during training process. Thus, this method improves model by taking two aspects: one is to improve the accuracy of network by removing noise data, and another one is to improve the efficiency of network by reducing the size of data.

In terms of regression problems, mark prediction is a major task, especially in the education system. Many students always behave badly in their final exams due to psychological factors and environmental factors, so they could not expose their real powers and even fail exams. Therefore, for schools, predicting students` final marks is important to reduce this risk, and for students, it can reduce their psychological stresses and facilitate the expression of their real powers. This paper is aimed to exam this prediction ability by applying BDR method in BP. Additionally, in this paper, genetic algorithm (GA) is adopted before training process for feature selection, so the performance of GA can be also examined in terms of accuracy improvement by optimized features. The purpose of feature selection is, by choosing a subset of all existing features, useful features can be identified and used to reduce the size of network and to improve the performance of network [3].

This paper mainly focuses on the following topics:

1. Predicting final marks as accurately as possible. Data set [6] is quite limited, and only a hundred instances with 14 features are provided, so it is impossible to improve its accuracy highly.
2. Performance analysis of BDR under the limited data set. Its purpose is mainly to discover some advantages and disadvantages of BDR.
3. Analysis of performance of GA applied for feature selection before training.

2 Method

In this part, detailed process from network design to parameter updating for two different networks is introduced for the purpose of both reproducibility and falsifiability.

2.1 Dataset Preprocessing

Preprocessing dataset in an appropriate way is a key aspect to ensure that the generalization of a network model is at a high level [2], and it is also difficult to be implemented when the dataset and its features are very large in quantity. In a large dataset, there are must be many noise data entities and useless features that cannot be identified easily, and a one-

size that fits all situations does not exist. Therefore, trying various preprocessing methods is a good way, even in a small-size data set.

The data set applied in this paper is quite small. In the data set, only 153 instances with 15 features and 1 target are provided, and plenty of values of marks are missing which are filled with '.', and these values are replaced by 0, assuming that students did not submit their works or take exams. Besides, the feature 'Regno' is unused in training process, as it is only an identifier of different students and would not be used to update model obviously, and the feature 'Tutgroup' contains some missing values, so it is also removed for convenience. Except 'Regno' and 'Tutgroup', the rest of features (13 features) are coded and used, as GA would be applied to differentiate useful features from them. Table 1 shows the result of preprocessing. Furthermore, the original data set is divided randomly into three parts: training set (60%), validation set (20%) and testing set (20%)

Feature	Preprocessing Result
Regno	[Unused]
Crse/Prog	1 – 23
S	1 and 2
ES	1, 2, and 3
Tutgroup	[Unused]
lab2	[Original]
tutass	[Original]
lab4	[Original]
h1	[Original]
h2	[Original]
lab7	[Original]
p1	[Original]
f1	[Original]
mid	[Original]
lab10	[Original]
final	1, 2, 3, and 4

Figure 1 Preprocessing Result

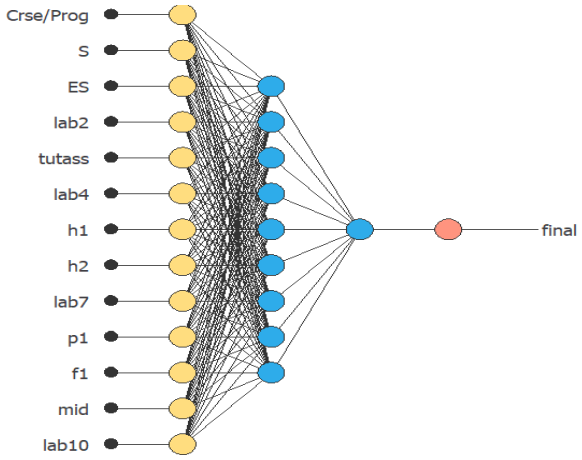

Figure 2 Network Structure

2.2 Network Design

A classification network is established to predict final, shown as figure 2. In the network, there are two fully connected layers with a hidden layer, and the activation function of the hidden layer uses Sigmoid function. Sigmoid function is mostly used as an activation function in feedforward neural network [4], and it has some advantages in this classification problem. Firstly, due to its smooth gradient, output values can keep changing smoothly, and any sudden instability can be avoided. Secondly, output values can be compressed into the interval (0,1) by normalization, so these values, as inputs, can impact on the next neuron at same degree. However, the computation of sigmoid is more expensive than others. Besides, sigmoid function is often used to enhance understanding in simple networks, such as this network [5]. As mentioned in 'Dataset Preprocessing' section, the number of inputs is 13, and the number of outputs is only 1. Further, it is suggested that the number of hidden neurons are two-thirds of input features [6], so the number of hidden neurons is set to 9 ($\approx 13 \times 2/3$).

2.3 Feature Selection – Genetic Algorithm

Before training process, Genetic Algorithm (GA) is adopted to select valuable features that can be used to train model. GA is a method that uses the theory of 'biological evolution' proposed by Darwin as a reference and globally searches potential solutions of problems [7]. The major steps of GA are shown in figure 3. The basic idea of GA is to map the search space to the genetic space and encode each possible solution into a vector – chromosome. Each element of the vector is called gene. By continuously calculating the fitness of each chromosome, the best chromosome is selected, and the best solution is obtained. It is worth emphasizing that because of GA`s nature – probabilistic method, the final solution could be not optimal [7]. GA is often used to address problems of combination optimization [8]. In combination problems, brute-force approach can be applied theoretically to find the best solution of combination, but it would take much time (2^n possible subsets if there are n features). Therefore, GA is a good method to address it.

2.3.1 Steps of GA in This Paper

To start with GA, its population needs to be created. As mentioned, GA is a probabilistic method, so the process should be random. In this dataset, there are 13 features used to train the model, so chromosome can be presented by binary string whose length is 13. For each binary string, if a character (or gene) is 1 it means that related feature is used for training, and if a character is 0 it means this feature is unused. In this paper, its population size is set to 200.

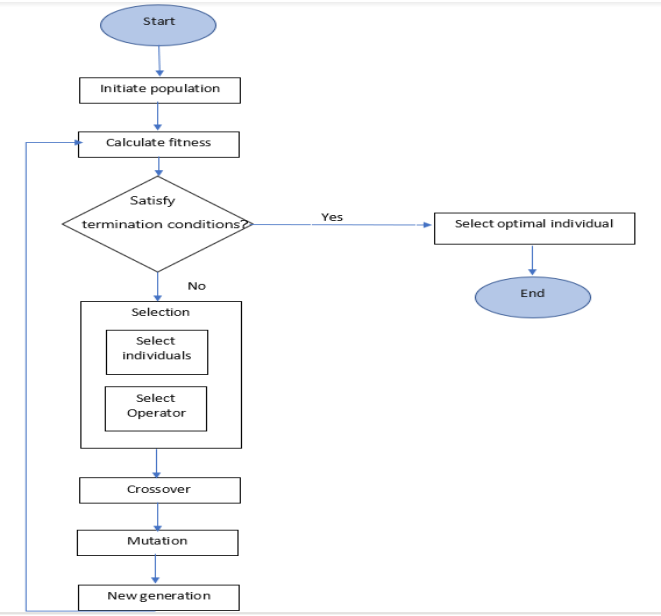

Figure 3 Steps of Genetic Algorithm

When population is created, fitness needs to be calculated for each chromosome. In the paper, the goal is to select a subset of features that minimize its mean squared error of cross validation. Fitness function would return sorting score list and related chromosome list, and these two lists will be used for the selection process.

After calculating fitness, 40 best chromosomes are selected based on score, so that the population moves towards global optimal solution, and the solution will not fall into a local optimal when another 40 chromosomes are selected randomly as well. In crossover process, two chromosomes are used to create the next generation. Parents are selected from the integration of 40 best chromosomes and 40 random chromosomes. In mutation process, small change of chromosome can avoid fast convergence to a local optimal, and this change involves the random exclusion of features with a small probability (5%). When a new generation is created, fitness would be recalculated until there are 10 generations.

2.4 Validation Method

To avoid overfitting and improve generalization of model, early stopping method are adopted. The specific method of early stopping is to calculate the accuracy of validation data at the end of each epoch, and training will be stopped when the accuracy is not improved in an epoch. However, because of limited validation set, it would not reflect the reality for sometimes [9]. Therefore, what did in this paper is to record the best validation accuracy so far during training, and when epochs fail to reach the best accuracy for 10 consecutive times, it can be considered that the accuracy will not improve any more, and training can be stop at this point.

2.5 Noise Reduction - Bimodal Distribution Removal

Comparing with the normal backpropagation algorithm, bimodal distribution removal (BDR) tries to detect and remove some noisy data when some conditions hold at each training iteration. Both efficiency and accuracy of network can be improved. The major steps of BDR are shown below:

1. Set an upper bound for the purpose to enable BDR.
2. Train the network with original training set in the way of normal backpropagation algorithm.
3. When obtaining error PE of each pattern at each iteration, calculate the variance of this error list.
4. If variance is smaller than the upper bound, enable BDR.
5. Calculate the average error AE of all training set.
6. If PE is greater than AE, put PE into a list.
7. Calculate the list`s average m and standard deviation sd.
8. If PE is greater than the sum of m and the n times of sd ($n \in [0,1]$), delete the pattern from training set permanently.
9. Update upper bound appropriately to enable BDR next time.

According to its steps, BDR has a problem. If there are too many patterns removed from dataset, the generalization level of the model could be reduced [4], so a stopper may be required when enough patterns are removed. In this paper, the validation method 'early stopping' are used to stop it when enough patterns are removed, and to avoid over-fitting.

3 Result

In this part, some major discoveries from actual network training and testing are illustrated and analyzed.

3.1 Error Comparison

After completing 50 epochs, the error distribution of networks with and without BDR are shown in Figure 4. When BDR is involved within the network, the error distribution is more centralized within a lower range of error. By contrast, the error distribution of network without BDR is more dispersed.

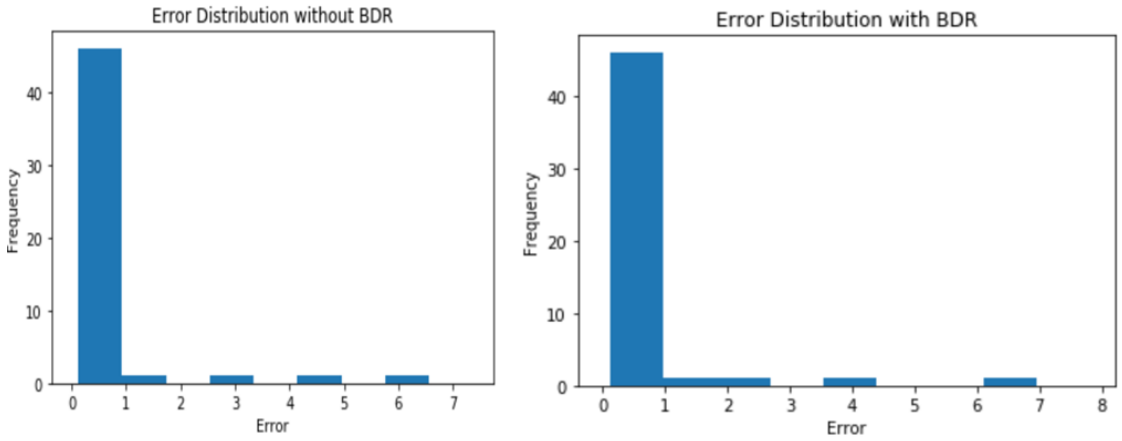

Figure 4 Comparison of Error Distribution

3.2 Accuracy and Loss Comparison

After completing about 600 echoes, early stopping is applied to avoid overfitting. Loss and accuracy of BP network with and without BDR is shown in Figure 5. It shows that BDR decreases training loss and improves training accuracy significantly, while testing accuracy is not improved much. It is initiative because the purpose of BDR is only to prevent overfitting and improve performance by removing information from training set, and it cannot affect testing dataset directly. Therefore, the improvement degree of testing actually depends on the similarity between training set and testing set. If their similarity is relatively high, testing accuracy can be improved significantly; otherwise, testing accuracy only can be impacted slightly.

Network	Training Loss	Training Accuracy	Testing Accuracy
BP without BDR	0.9770	38%	36%
BP with BDR	0.2117	71%	40%
BDR + GA	0.2461	56%	45%
GA	0.2128	70%	33%

Figure 5 Comparison of performance of network with and without BDR

3.3 Genetic Algorithm Evaluation

Cross validation score is decreased when GA generation is increased from 0 to 10 in this dataset, shown as Figure 6, and it demonstrates that by removing some unrelate features, shown in Figure 7, over-fitting can be reduced to some extent. Because the size of original dataset is limited, the performance of the network can be improved by deleting some data and features.

Besides, as shown in figure 5, the training accuracy of network with both BDR and GA is lower that network with GA only, while the testing accuracy is just the opposite. It demonstrates that removal of both features and patterns can improve the performance of network, especially for testing dataset. However, from figure 5, GA can decrease high training accuracy caused by BDR. As a bold assumption, it is probably because information removed by BDR and GA at the same time may include some useful information, and it would slightly decrease the performance of network.

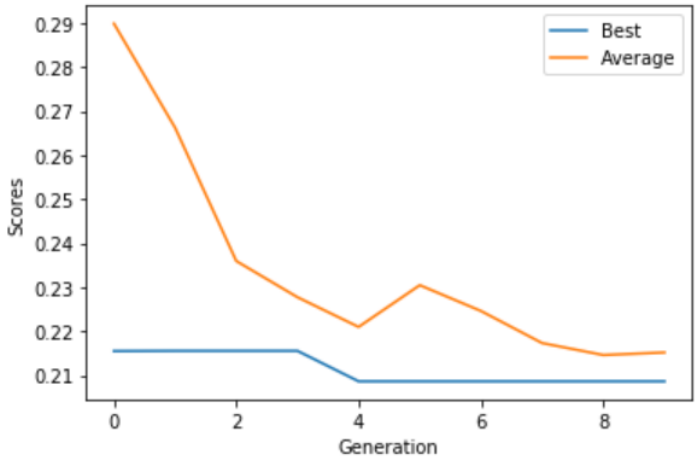

Figure 6 Cross Validation Score and Generation

Features	Cres	S	ES	Lab2	tut	Lab4	H1	H2	Lab7	P1	F1	mid	Lab10
Gene	1	0	1	0	0	0	1	1	1	1	1	1	0

Figure 7 Final Feature Selection

4 Conclusion

According to above results, by involving the BDR within BP network, the performance of testing set can be improved significantly, especially for training dataset. The degree of improvement of testing dataset may depend on the similarity between training dataset and testing dataset. In addition, application of GA can improve model generalization by removing additional features, but the combination of GA and BDR can decrease training accuracy.

For future work, the reason why the combination of GA and BDR can decrease training accuracy can be explored for further step.

.

References

1. Fu, B, Zhao, X, Li, Y, Wang, X & Ren, Y 2019, 'A Convolutional Neural Networks Denoising Approach for Salt and Pepper Noise', Multimed Tools Appl, vol. 78, no. 1, pp. 30707-30721, viewed 4 May 2020, https://arxiv.org/ftp/arxiv/papers/1807/1807.08176.pdf

2. Kotsiantis, S, Kanellopoulos, D & Pintelas, PE 2006, 'Data Preprocessing for Supervised Learning', International Journal of Computer Science, vol. 1, no. 1, pp. 111-117, viewed 2 May 2020, https://www.researchgate.net/publication/228084519_Data_Preprocessing_for_Supervised_ Lea rning

4. Slade, P & Gedeon, TD 1993, 'Bimodal Distribution Removal', Lecture Notes in Computer Science, vol. 686, no.1, pp. 249-254, viewed 31 March 2020, https://doi.org/10.1007/3-540-56798-4_155

5. Li, X n.d., 'Applying Genetic Algorithm and Bimodal distribution removal to improve classification problem', Semantic Scholar, viewed 29 March 2020, https://pdfs.semanticscholar.org/21db/14631d0974c4559108d4a2e3d53b0d2e7db4.pdf

6. Che, E, Choi, Y & Gedeon TD 1995, 'Comparison of Extracted Rules from Multiple Networks', IEEE, vol. 95, pp. 1812-1815