Experiment of Bimodal Distribution Removal on Medical Dataset

Haowen Li

Research School of Computer Science,
Australian National University
Email: u6342101@anu.edu.au

Abstract. How to clean up noisy training sets is always a hot spot in machine learning, especially for medical dataset. A lot of methods about removing outliers have been created during past decades. This report focuses on a kind of statistically based method, Bimodal Distribution Removal, and tests it on a hypothyroid diagnosis dataset. The experiment bases on a two-layer fully connected network with feature selection achieved by genetic algorithm. This paper finds this method is hard to distinguish uncommon data points from noisy and halt the training at inopportune moment, which decline the performance of neural network on test dataset.

Keywords: Neural networks, noisy data, outlier detection, unbalanced data, halt condition

1 Introduction

Nowadays, the computer-aided diagnosis expert system plays more and more significant role in medical procedure. But there are still many problems need to be solved: the medical datasets are always unbalanced, which decreases the performance of predicting model; the diagnosis process always involving a lot of indicators, which extends the training time as well as the size of neural network; the doctor cannot ensure every diagnosis is perfect, which means the training set may contain outliers or incorrect data. In order to tackle these problems, this paper uses special sampling method to balance the dataset and genetic algorithm to select useful features. But the main point of the paper is thoroughly investigating the utilize of an outlier detection method in disease diagnosis.

The method tested in this experiment is Bimodal Distribution Removal (BDR) [1], which is a kind of outlier detection method. It works by removing the pattern with high error during training the network. However, in case of unbalance dataset, the rare data also can result in high error in the early period of training. It is vital that whether BDR can distinguish rare data or meaningless outlier.

This experiment was performed on a hypothyroid dataset [2]. The dataset has 21 attributes and 3 classes. The task of this dataset is using the attributes which are gathered from medical tests to determine two kinds of hypothyroid, primary hypothyroid as well as compensated hypothyroid, from healthy cases. The whole dataset is divided as training dataset (3772 cases) and testing dataset (3428 cases). Because there are 92 percent of the cases are not hypothyroid, the cases of hypothyroid are possible to be defined as noisy by outlier detection methods. That is one of the most important reasons why this dataset was chosen.

2 Methods

2.1 Evolutionary feature selection

The epoch of big data brings massive information to researchers, which promotes the development of machine learning. However, a large number of features also can become a kind of problem. Redundant features not only result in the waste of the time and space used by neural networks, but also weaken the predication performance of networks. Therefore, a feature selection method based on genetic algorithm was proposed. This method regards the selection of features as chromosome. For example, the chromosome '01' means the first feature is not used in training and the second feature is used. Then, using the selected features to train the neural network and the fitness is the network's classification performance. After repeating the training as well as the selecting of these features subsets for several times, using the features subset which has best performance in last round in the formal training.

In this experiment, the population is 10 and the number of generation is 20. Each evaluation trains the network by 200 epochs. Lastly, the number of features decreases from 21 to 13 after using genetic algorithm to select features.

2.2 Sampling technology

The sampling method used by this paper is simple and intuitive. Firstly, classifying all patterns into three categories according to their labels. Then duplicating a kind of rare patterns until their proportion exceeds a certain number (30% etc.) and mixing these duplicated patterns and common patterns. Repeating above operation for the other kind of rare

patterns. At last, randomly selecting some patterns from the mixing dataset as balanced dataset and the size of this dataset should be similar to the size of original dataset.

In this experiment, the proportion of disease change from (hyperfunction : subnormal functioning : normal) = (0.03 : 0.05 : 0.92) (original) to about (0.2 : 0.3 : 0.5).

2.3 Bimodal Distribution Removal

Bimodal Distribution Removal (BDR) is a kind of statistically based outlier detection method. BDR assume that in the very early period of training, the erroneous patterns would result in much higher error than other normal patterns. Therefore, the bimodal distribution is formed in the error distribution of early training. One big peak containing low error patterns that are learnt well by the network, and the smaller peak containing high error patterns that are outliers. The patterns between these two areas should be some rare patterns. They are useful and need to be slowly learnt by the network.

This method attempts to find the erroneous patterns by their relatively higher error, and then remove them step by step. To achieve this goal, firstly selecting patterns with error greater than the mean of all error into a subset. This subset must contain all of the outliers, because the number of outlier is less than useful patterns. However, this subset would also contain the rare patterns, and a part of normal patterns (if the number of normal patterns is largely greater than outlier). They are useful and should not be removed. So just removing the patterns with error greater than the sum of the mean and the standard deviation of subset errors. In addition, the value of standard deviation in last operation should multiply a discount coefficient between zero to one. Above operations should be repeated in every 50 epochs until the variance of all error less than a constant. At that time, the training need to be halted. Because the low variance signifies the smaller peak has disappeared and the network is trained well.

2.4 Neural network architecture

The algorithm was used is normal backpropagation and the fully connected network's structure is 13-20-3 for original dataset and 13-20-3 for balanced dataset. The number of input neurons is determined by the evolutionary algorithm and the number of hidden neurons is selected by several experiments with different hidden neurons. Each training for network without BDR continue 2000 epochs since that produce relatively better result than the other trainings with less or more epochs. When training the network with BDR method, the training would be halted by BDR method which uses the variance of error to judge whether the network is trained well or not.

2.5 Evaluation method

There are two kinds of comparation. First one is the comparation between the network using 21 features and the network using 13 features which is the results of evolutionary feature selection, which is about feature selection. Second one is the comparation between the network training without BDR method and the network training with BDR method, which focus on BDR method.

Both of two kinds of comparation using 3 indicators to investigate the performance of the network. The indicators are the accuracy of the network on test dataset, the diagnostic rate of patients (the number of patients who are labelled as any kinds of hypothyroid divided by the number of all patients) as well as the mean F1 score. They can well reflect the performance of the network in practice. The difference is that the first kind of comparation considers the time is taken while training the network. Because saving time is the prime purpose of features selection.

Each indicator is gathered by 5 times of training to prevent the fortuity. The time data are collected in the author' personal hardware (only for comparation).

3 Results and Discussion

First part is about evolutionary feature selection. There is a comparation between 4 networks with different number of features. Then, the results of training without BDR as well as training with BDR would be compared. At last, the comparation would be repeated upon a balanced dataset which comes from the result of the above sampling method.

3.1 Feature selection

Before using genetic algorithm to choosing features for the networks used to make comparation, we need to ensure this method indeed improves the performance and know how much the advance it can make.

Table 1. Performances of networks with different features on testing dataset

	Original dataset with all features (21 features)	Original dataset with selected features (13 features)	Balanced dataset with all features (21 features)	Balanced dataset with selected features (13 features)

Accuracy	0.969	0.974	0.942	0.930
Diagnostic rate	0.784	0.928	0.972	0.984
Mean F1 score	0.492	0.534	0.423	0.401
Training time	4min 51s	4min 43s	5min 48s	5min 29s

We can find that both the performance and training time are improved by evolutionary feature selection. Although the improvement on some indicators is not obvious, this method can be used for the following training with confidence.

3.2 Testing BDR method on original dataset

Before using the BDR method, the error distribution of early period of training need to be inspected. Only if the bimodal distribution exists, the BDR method can work.

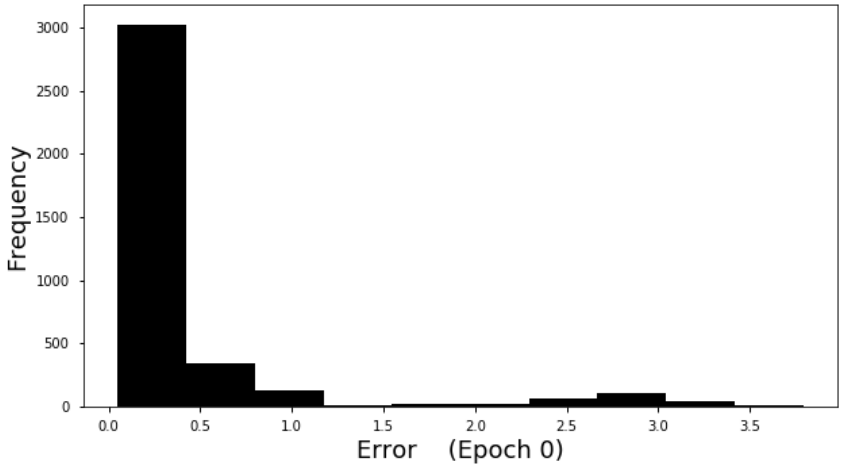

Fig. 1. Error distribution of original dataset at epoch 0

The error distribution is divided into two parts. Major part locates in lower error area and the smaller one contains some high error pattern. There is a gap between these two sections. All of these are like what BDR methods predicts.

However, the original dataset is unbalanced. We cannot make sure the high-error patterns are outlier or rare data point, or both of them. If after using BDR methods to remove these high-error patterns step by step, the performance of network on testing dataset became better, we could say the BDR method can distinguish between incorrect patterns and rare patterns.

Next step is comparing the performance of these neural networks.

Table 2. Performances of networks with different outlier detection methods on testing dataset

	Original dataset without BDR (2000 epochs)	Original dataset with BDR (325 epochs)	Original dataset without BDR (325 epochs)
Accuracy	0.974	0.936	0.943
Diagnostic rate	0.928	0.184	0.380
Mean F1 score	0.534	0.323	0.338

There is a performance decline after applying BDR method to remove outliers. Noticing that the diagnostic rate is only 0.184, which means no more than 4 out of 5 real patients would not be diagnosed as hypothyroid.

One thing need to be pointed out is BDR method halt the training in much earlier period compared with normal situation. In most cases, the training was halted just after 325 epochs when BDR method permanently remove the fourth batch of patterns. But for the training without BDR method, the performance on testing set would be improved by extending training period from 1000 epochs to 2000 epochs, which means BDR method halted training before the network learnt well. Although that situation can be improved by changing the value of BDR method's hyperparameter, it is difficult to extend the training time with BDR method to the enough length.

Another theory can be used to explain the decreasing of performance. It is possible that the high-error part contains value patterns and the BDR method remove these rare data points which are the key portion of this problem. Therefore, the loss of patients' data lead to the network cannot detect the disease.

To confirm the main reason of the declined performance, the result of 325-epochs training without BDR was introduced into table. Its performance is a bit better than the performance with BDR method (seeing Table 1). Then we know BDR method remove a part of useful data in deed but that is not the main reason. What should be blamed is the prematurely stopped training.

To improve the performance of BDR method, the time of training must be extended. Changing the value of BDR method's hyperparameter may slow down the shrinking of variance, then the training would not be halted prematurely.

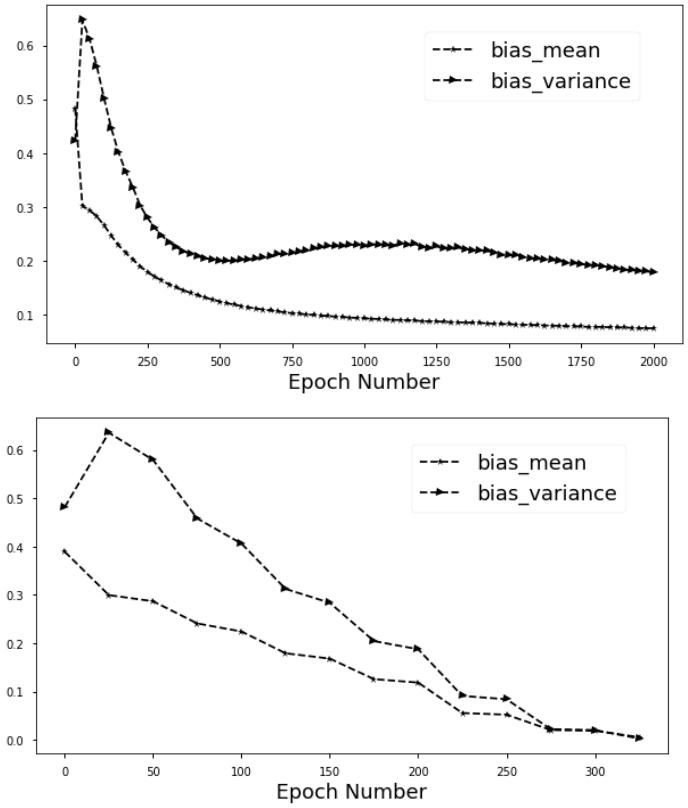

Fig. 2. Change of mean error and variance of error during training on original dataset. First plot belongs to training without BDR method. Second plot belongs to training with BDR method.

The changes of variance during trainings with two conditions are very different. For training without BDR method, the variance of error decline significantly in early period. Until it reached 0.3, the trend of declining slow down. But in case of training with BDR method, the downward trend of variance did not change until the value of variance becomes very small and training was halted at that time. Several ways try to slow down the declining are all failed. Once some high-error patterns are removed, the variance would inevitably drop.

In the experiment, it has been certified that BDR method is not suitable for this dataset. The major problem is BDR would halt the training in very early period.

3.2 Testing BDR method on balanced dataset

BDR method remove the useful patterns in the last test because the patterns are so rare that are determined as outlier. Would BDR method perform better in case of balanced dataset? In this part, the networks are trained by a balanced dataset. The dataset is made by repeating the patterns of hypothyroid and randomly selecting a part of normal patterns to maintain the size of training set.

First thing should be done is inspecting the error distribution at epoch 0.

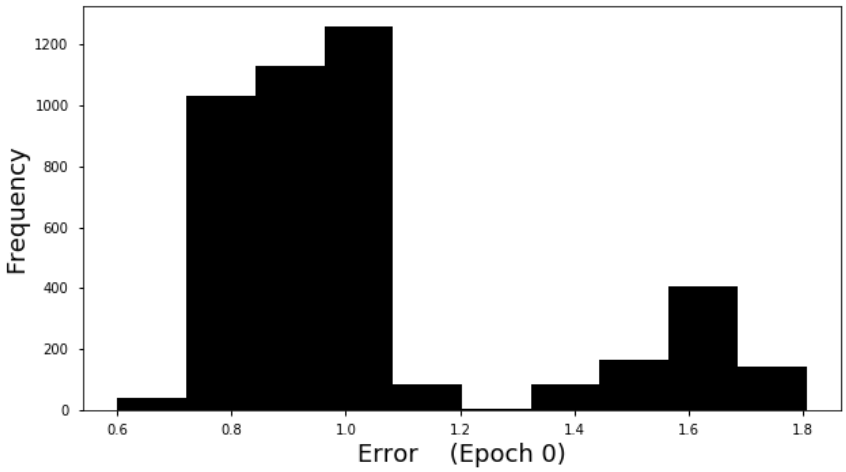

Fig. 3. Error distribution of balanced dataset at epoch 0

Similar to the situation of original dataset, the error distribution is bimodal. But the difference is the number of high-error patterns is greater than the last one (see Fig 1), which should not happen.

If the dataset is balance, the proportion of hypothyroid's pattern would much higher than before. Therefore, these patterns should not be treated as outliers by the network and should not produce high error as before. There are two possible explanations. First one is in the process of sampling, more noisy patterns were included into balanced training dataset. Secondly, there are still a part of hypothyroid's patterns which are recognized as noisy data point. Noticing that not all patients' patterns would be in that area, because these patterns take up half of training set. For the first possible explanation, the sampling upon training set was repeated several times. In each time, the similar situation appeared. Although the sizes of high-error part are not exactly as same as last time's, all of them greater than the same part in Fig 1. So, the first one is not reasonable, and second explanation becomes more possible.

Then the same test of these new trained networks is carried out on same testing set. Table 2 also includes the results reported by Weiss and Kapouleas [3], which only contains the accuracy of test set. Their solutions are rule-based. I think that is the main reason why their classifiers' performances are much better than the networks were trained in this experiment.

Table 2. Performances of networks with different outlier detection methods on testing dataset

	Balanced dataset Without BDR (2000 epochs)	Balanced dataset with BDR (375 epochs)	Balanced dataset without BDR (375 epochs)	Weiss' 21-3-3 Neural Network (2000 epochs)
Accuracy	0.930	0.828	0.932	0.985
Diagnostic rate	0.984	0.716	0.952	-
Mean F1 score	0.401	0.266	0.400	-

After adjusting the proportion of 3 classes pattern in training dataset, BDR method would halt the training in epoch 375. Just like what is done in Table 2, the result of 375-epochs training without BDR method was added into comparison.

Surprisingly, while the performance of 375-epochs training without BDR method almost equal to 2000-epochs', the performance of training with BDR on balanced dataset still is the worst one. It is obvious that its bad performance cannot blame for halting the training of underfit network. According to the former assumption for Fig 3, now we can ensure the main reason should be BDR method remove some useful data points from training set. We also can assume BDR also remove these data in the training for original data. But these patterns just occupy small proportion of original data, which reduces the impact of removing these patterns.

4 Conclusion

This experiment found that for selected real-world disease dataset, the application of Bimodal Distribution Removal in neural network's training gathers unsatisfactory performance. The major problem is BDR method rely on network itself to detect the outliers and control the halting condition of training.

In case of unbalanced training set, BDR method stop the training when network is still underfit. The problem is this method remove the patterns which result in relatively high error, then the variance of all errors would decrease rapidly. It is the variance that the method uses as halting conditions. Therefore, the premature halting would always happen, then only the underfitting network can be produced by the training with BDR method.

In case of balanced training set, many useful patterns also produce high error. But BDR method fail to distinguish these patterns from noisy. After this method removing the patterns belong to patients, the network cannot recognize these patterns, then the diagnostic rate significantly decreases.

In conclusion, for some kinds of dataset, BDR method would halt the training in very early period, which produces underfit network. In both unbalanced and balanced dataset, BDR method would remove high-error patterns, no matter whether it is outlier. There is much room for improving this method.

5 Future Work

The experiment only tested single outlier detection method on two versions of a kind of dataset. There still are a lot of thing can be done for future research, including:

- More experiments on other datasets and compare results with other outlier detection methods.
- Change the algorithm of BDR method to improve its performance on similar datasets like this disease data. Both the threshold of removing patterns and the halting condition can be good point.
- Utilizing rule-based method on the training of neural network, which would help us understanding this method's influence on BDR method.

- The results of this experiment related to some hyperparameters about training neural network, such as learning rate, the number of hidden units or hidden layers, and the methods of sampling or updating the weights. It is possible that the network distinguishes useful data from outliers in some cases.

References

1. Slade P., Gedeon T.D.: Bimodal distribution removal. In: Mira J., Cabestany J., Prieto A. (eds) New Trends in Neural Computation. IWANN 1993. LNCS, vol 686. Springer, Heidelberg (1993)
2. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml
3. Weiss, S.M. and Kapouleas, I.: An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In: Sridharan N.S. (eds) Proceedings of the 11th international joint conference on Artificial Intelligence, vol. 1, pp. 781--787. Morgan Kaufmann, San Francisco (1989)