Training Stage Optimization for Neural Network Operating On Small Scale Data Set

YILIN LI[1]

[1] Australian National University u6508701@anu.edu.au

Abstract. Neural Networks is a mathematical model commonly used in data prediction. It is widely used in target detection, face recognition and behavior recognition. The noise data has a great influence on the training of the neural network, in the back-propagation step, the amount of change of each weight depends on the difference between the network output .When presented to the network, the error pattern in the training set can make a big difference between the expected and actual output. Noise data (outlier) in nature word is very difficult to find. In addition, the method of optimizing the parameters of the neural network is also very important for training. This paper use data to predict which language level (L1 or L2) the participants will belong(classification) and what is the total score will get(regression) to test Bimodal Distribution Removal algorithm (BDR) and an improved method called Isolated Forest Removal (IFR)is proposed, Genetic algorithms (GA) are also used to train the parameters of neural networks. The result show that the Genetic Algorithm (GA) has a good performance in optimizing the parameters of the neural network. The result also shows that the limitation of the Bimodal Distribution Removal algorithm on small scale data set is sensitive for threshold and parameter, so its stability is not as good as isolated forest removal. The isolated forest removal algorithm can control the proportion of abnormal data by changing the threshold, and its performance is better. The Genetic Algorithm can enhance both two algorithms' performance and even overcome the drawback of Bimodal Distribution Removal algorithm .Therefore, on the task of small-scale data set ISR has more advantages and after applying Genetic Algorithm (GA) for training stage, both two algorithm can have good performance on small scale dataset. The resulting system show that neural networks performance on small scale data set may be different from its on enough data set.

Keywords: Regression, Classification, Neural Networks, Optimization, Genetic algorithms, Bimodal Distribution Removal Isolated Forest Removal

1 Introduction

Today's digital environment allows people to access large amounts of information quickly and easily. We are being forced to accept a lot of information that we do not need. However, at any time Everything has two sides. Visual interference usually exists in the form of web pop-ups, video advertisements and distract users. Studies have shown that auditory disturbances (such as background noise) can impair reading comprehension[1]. This raises the question of whether visual interference will negatively affect the reader. This research question aroused my interest, which is reason I chose this dataset.

According to paper [2], data were collected from 22 participants (8 women) with an average age of 21.7 years (standard deviation 3.0). All participants had normal or corrected vision. Participants were mainly (N = 18) recruited from the first year of computer science courses on web development and design offered by the Australian National University (COMP1710).

The remaining participants are students from the Australian National University. Participants are divided into two groups. Those who mother language is English label as L1, and those who second learned to English is second language, L2. There are 4 L1 participants and 8 L2 participants.

The experiment time is about 30 minutes. First, explain the experiment to the participants, and then ask them to read and sign the consent form. Participants received a pre-experiment questionnaire. Calibrate the EyeTribe eye tracker until you get a "perfect" calibration based on the tracker.

After calibrating the routine, the participants read the text while monitoring and recording their sight. Then, the participants answered 10 understanding questions about the text. Regardless of reading level, comprehension questions are always the same. Finally, provide participants with a post-experiment questionnaire.

The data like Number of fixations、page total and per region、Total fixation duration (seconds) 、Num fixations out (in) of text area、Reading ratiowas was recorded.

This paper mainly uses data to make statistical inferences, so it does not involve too many classification and regression problems. The task for this report is to predict which language level (L1 or L2) the participants will belong(classification) and what is the total score will get(regression) using the data.

1.1 Genetic Algorithm (GA)

Genetic Algorithm (GA) was first proposed by John Holland of the United States in the 1970s[3]. The algorithm was designed and proposed according to the evolutionary laws of organisms in nature. It is a calculation model of biological evolution process that simulates the natural selection and genetic mechanism of Darwin's biological evolution theory, and is a method to search for the optimal solution by simulating natural evolution process. The algorithm mathematically uses computer simulation operations to transform the problem-solving process into a process similar to the crossover and

mutation of chromosome genes in biological evolution. When solving more complex combinatorial optimization problems, relative to some conventional optimization algorithms, it is usually possible to obtain better optimization results faster. Genetic algorithms have been widely used in the fields of combinatorial optimization, machine learning, signal processing, adaptive control and artificial life. This paper will use Genetic Algorithm (GA) to optimize the parameters of the neural network[3].

1.2 Bimodal distribution Removal algorithm and its limitation

In the back-propagation step, the amount of change of each weight depends on the difference between the actual network output and the desired output. When presented to the network, the error pattern in the training set can make a big difference between the expected and actual output. When the network tries to minimize the errors of these modes, this will produce a larger weight change. Since most networks are trained, these error patterns will not extend the training time until the mean square error in the training set is below a certain threshold. This increase in training time greatly increases the chance that the network is too suitable for the training set. To avoid this happen, Bimodal distribution removal (BDR) are design to solve this. BDR can work by learning the mean variance and standard deviation to find the outliers, research believe it have some advantage compare to other outlier detection method like, the network can identified the outliers during the training and outliers are removed during the training rather than before it, so the neural networks have enough chance to learn from the data. However, this research finds that BDR performance not good on a small data set[4][5]:

1. without enough data set BDR can hardly predict outliers during the training
2. BDR only take error(loss) into consideration, while sometime neural networks have learned from the outliers and its error will not be high, so BDR cannot find it.
3. BDR is sensitive for threshold and parameter.

1.3 Isolated Forest

Isolated Forest(iForest) is a anomaly detection method based on Ensemble. It has linear time complexity and high accuracy. It is a state-of-the-art algorithm that meets the requirements of machine learning or neural networks.
iForest is suitable for anomaly detection of continuous data. Anomalies are defined as "outliers that are easily isolated", which can be understood as points that are sparsely distributed and far away from high density data. Consider with data space, sparsely distributed areas indicate that the probability of data occurring in this area is very low, so it can be considered that the data falling in these areas is abnormal.
Assuming that a random hyperplane is used to cut the data space, two subspaces can be generated at a time. After that, we continue to use a random hyperplane to cut each subspace, and loop until there is only one data point in each subspace. Intuitively, we can find that clusters with high density can be cut many times before stopping cutting, but those points with low density can easily stop into a subspace very early[6].
Since the cutting is random, it is necessary to use the ensemble method to obtain a convergence value (Monte Carlo method), that is, repeatedly cutting from the beginning, and then averaging the results of each cutting. iForest consists of t isolated trees of iTree, each iTree is a binary tree structure[7].
For a training data x, we let it traverse each iTree, and then calculate how many layers x finally falls on each tree (x at the height of the tree). Then we can get the average height of x in each tree. After obtaining the average height value of each test data, we can set a threshold (boundary value), and the test data whose average height value is lower than this threshold value is abnormal.
Compare to BDR, Isolated Forest have following advantage:

1. Randomness: each isolated tree randomly selects some samples.
2. No distance or density parameter.
3. Linear time complexity.

1.4 Aim and Motivation

Maintaining attention is very important for daily life and studying. This research background makes me feel very meaningful, which is why I chose this data set. Out of the large research efforts devoted to Outlier detection algorithm before training, little research has considered the predict the outlier data when training, which are hard work to do. This work targets to check:

1. Do neural networks build meet regression or classification problems?
2. how Bimodal distribution removal will influence the neural networks during the training in a small scale data set and will it improve the result of training.
3. Compare to Bimodal distribution Removal, how Isolated Forest Removal performance.

2 Method

This task method includes the following parts: data loading and preprocessing, defining neural network model, model training and testing, and removal algorithm improvement.

2.1 data load and preprocessing

First read the xls file containing the data. It is worth noting that there are several sheets in the file. The required data is in the sheet named "for spss". The xls file is read using python pandas, and the time data in a special format is converted into seconds, and the minmaxscaler method is used to convert all data to 0-1 interval to improve the training effect and prediction accuracy. The dimension of the data is 66 * 21, the dimension of a single data is 21, a total of 66. 52 (80%) sample are split to training set while rest of 11 are made of test set.

2.2 defining neural network model

For the neural network of the regression task, the number of neurons in the input layer is the same as the feature dimension of the training set, and the number of hidden layers is 2. The reason for this setting is that, according to research, it is not suitable to use too Complex model. A simple model can avoid the problem of overfitting. The setting of the output layer is the number of prediction categories of the classification problem 2. The setting of the learning rate is determined by the parameter adjustment. First, select a smaller learning rate to initialize, and then continue to increase. When the loss function does not converge, the number of training needs to be increased. Finally, ensure that the learning rate and the number of training times are at a reasonable value, so as to avoid too much learning rate or too long training time and waste time. The difference to the regression problem is that the output layer has only one neuron and does not need to be processed by the activation function.

2.3 define Genetic algorithms

Chromosome is a list with three elements, the first element is the activation function in the classification problem, the number of hidden layers of the second element, and the third is the learning rate. The initial population is 10, use Two point crossover, the first crosspoint is between the first gene and the second gene, and the second crosspoint is between the second gene and the third gene. Fitness Function is the performance of the neural network with parameter chromosome information on the test set. For classification problems, it is the accuracy of the test set. Regression problems are the mean square error . Two of the best performing offspring can be added to the population.

2.3 define Bimodal distribution Removal algorithm

First, start training. During the training process, use NumPy to define a matrix to record the error of a single training data during the training process. The number of rows in the matrix is the same as the number of data sets, and the results of the column and the prediction output are the same. loss. Then calculate the variance v_{ts} of all training data sets loss. The next step is to calculate mean error $\bar{\delta}_{ts}$, which can be obtained by summing and then averaging the NumPy. Sum function. After that take from the training set those patterns error greater than $\bar{\delta}_{ts}$, in order to do so, use a new numpy array to store these patterns for later steps. Calculate the mean $\bar{\delta}_{ss}$, and standard deviation σ_{ss} of this skewed subset, call numpy standard deviation function to get σ_{ss}. Permanently remove all patterns from the training set which error great than $\bar{\delta}_{ss} + \alpha\sigma_{ss}(0 \leq \alpha \leq 1)$, Repeat previous steps every 50 epochs, until normalised variance of errors over the training set $v_{ts} \leq 1$. Finally, encapsulate the above steps into a function, input The training set and test set also have a loss matrix that records the training set, and outputs the updated test set

However, some problems were found in the process of applying this algorithm. The first reason is that the data set is too small. If the data is removed once every 50 cycles, then a large part of the training set may be deleted. In addition, setting an appropriate v_{ts} needs to consider the speed of the original training first, because the Bimodal distribution Removal (BDR) algorithm has not completely learned the data at the beginning of the training, and the error variance is very large at this time, so there is a certain probability It will erroneously judge normal training data as noisy data. Finally, when judging the removal of the data set, the α in $\bar{\delta}_{ss} + \alpha\sigma_{ss}(0 \leq \alpha \leq 1)$, also needs to be properly set, so by observing the initial neural network loss function change curve, when the number of iterations can be considered that v_{ts} is stable instead of 0.1 set in the literature, the BDR algorithm is used to detect abnormal data once. When several training sets are removed, the overall training error may change significantly, and should wait long enough and after the number of iterations the next removal can be started. An example of failure in the study is shown in Figure 1. Because the data set

is too small and the α setting, half of the training data set was deleted by mistake, and the variance of error also fluctuated greatly. The test results are also very poor.

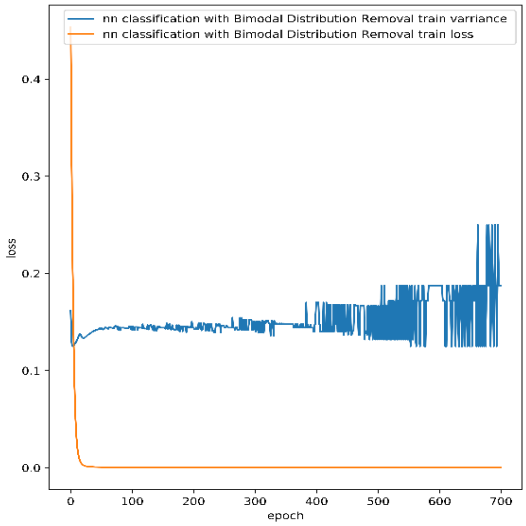

Fig. 1. A failure example of Bimodal distribution Removal example

2.3 modified BDR using Isolated Forest

Similar to the Bimodal distribution Removal (BDR) algorithm, the error matrix of a single training set is calculated first. Unlike the BDR algorithm, the isolated forest does not simply detect the loss feature, so it is necessary to add a training to the original training set The dimension of error, on this basis, uses the isolated forest algorithm to detect abnormal data. Then call IsolationForest in sklearn.ensemb to train random_state is the seed used by the random number generator, settled as 0. contamination is the proportion of outliers in the data set, settled as 0.05 and 0.01 for classification and regression, respectively. n_estimators is the number of base estimators in the ensemble, default as 100. Then we get the predict label for those training data and which label is -1 will treated as noise data. Finally, the above steps are encapsulated into a function. The input training set and test set also have the loss matrix recording the training set and the feature matrix with training errors, and the updated test data set is output.

3 Result

The number of iterations of the Genetic Algorithm will affect the accuracy of the results, but excessive number of iterations is very time-consuming. Table1 shows the performance of the optimized neural network under different iterations on the test set. 10 iterations take about two times while 5 and 20 takes 30s and 5 minutes respectively, so the number of iterations of the genetic algorithm is set to 10 after considering the overall accuracy and time consumption.

Table 1. Analysis of epoch of GA

Method	Unit	5 Epoch	10 Epoch	20 Epoch
Classification	Accuracy	71.42	71.42	64.28
Regression	MSE	0.0325	0.0227	0.0116

The mutation probability of the genetic algorithm will affect the training effect. When the mutation probability is too low, excellent offspring cannot be obtained, and the high mutation rate will make the excellent genetic information unable to be retained. Table2 shows the optimized neural network under different mutation probability Performance, so the mutation in future training is set to 0.1

Table 2. Analysis of Mutation rate

Method	Unit	Mutation 0.1	Mutation 0.2	Mutation 0.5
Classification	Accuracy	78.51	64.28	64.28
Regression	MSE	0.0072	0.0268	0.0301

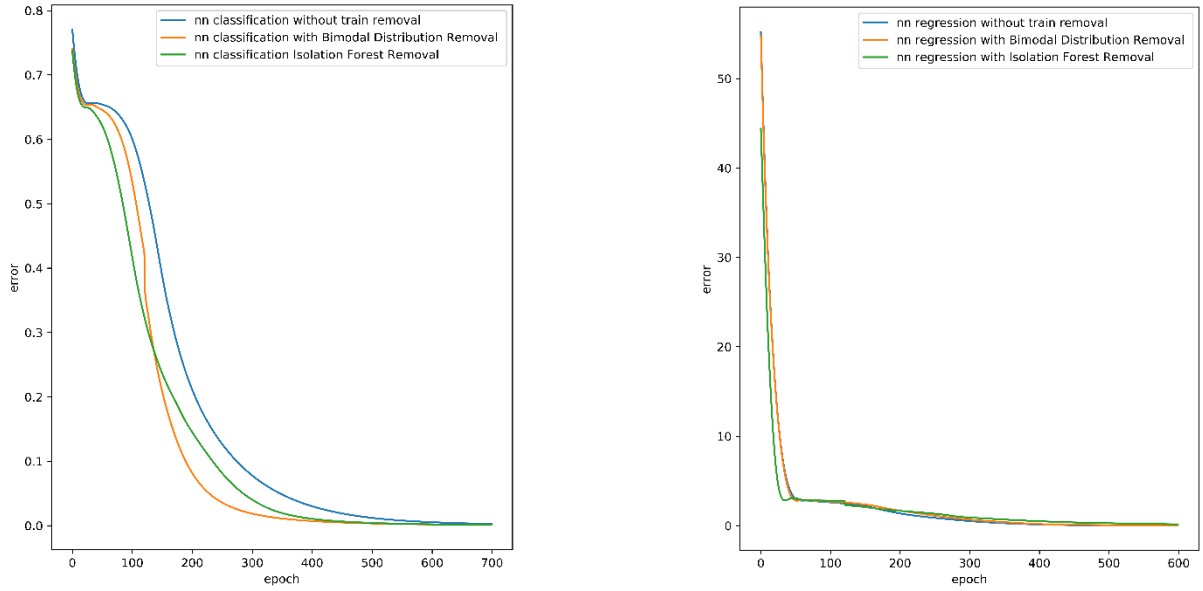

Fig. 2. The comparison loss changes during training figure of classification task(left) and regression task (left)for three different method

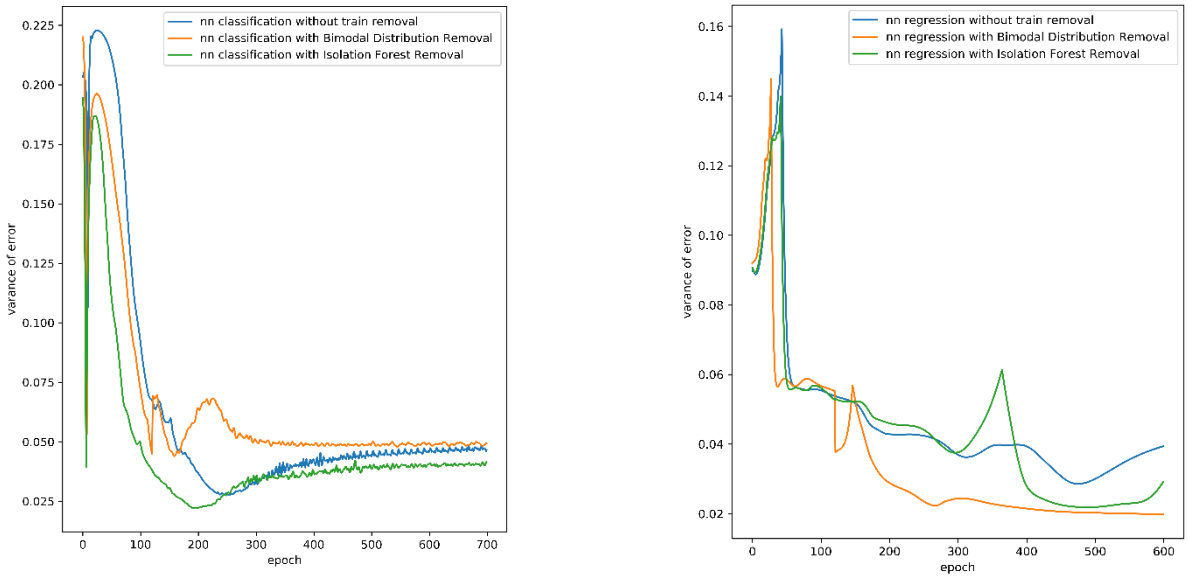

Fig. 3 The comparison variance change figure of classification task(left) and regression task(left) for three different method

From picture two and three, we can observe that in the classification problem, the loss of training loss without any algorithm is slower than the use of both, which is consistent with the research results in the previous paper. In the regression problem, the training decline speed of the three is basically the same, there is not much difference. Considering that the training data set in this study is small, using the two removal algorithm on a larger data set will improve the training speed. It can be found from picture two that when training reaches 120 steps, the loss variance of the two removal algorithms has changed significantly, but the effect of this change on subsequent training is undeterminable, and may be better or worse than without removal. The algorithm is better or worse. In the classification training, the three variance curves finally stabilized, and the variance in the regression training did not show obvious signs of convergence as the training progressed.

From Table 3 first 5 column, it is not difficult to see from the test results of the classification problem that the models of the three different methods all have a certain overfitting, that is, the performance on the training set is acceptable, but the performance on the test data set is not good. In the prediction, the neural network model tends to classify all the test data into the same category, so it can be concluded that the neural network model used in this study did not successfully learn to testers who predict different capabilities from the data set. For comparison, a random forest classifier was also added in this study, which also failed to predict the category of the test set, so it can be basically concluded that the reason for the poor classification effect is not related to the setting of the neural network model. The isolated forest removal algorithm was used to remove three data sets that were considered abnormal during training, but the training process using the BMR algorithm was not removed. This may be because the threshold of the BMR algorithm intervention was not set properly, resulting in the BMR algorithm not The coefficient setting for operation or standard deviation is not adequate. Through repeated experiments, we can find that the effect of the two removal algorithms on the results is unstable, sometimes it will improve and sometimes reduce the accuracy of the test.

Unlike the classification problem, the neural network model is excellent in predicting the test score, whether it is the test set or the data set, and the effect of the isolated forest removal algorithm is better than the other two. The root-mean-square error of the prediction results is less than about 1, and the training convergence speed is very fast. In the isolated forest, one outlier was removed during training, and the mean square error of test l dropped from 0.29 to 0.21, while the two outliers were removed by the BMR algorithm. It verifies the previous conjecture that the performance of isolated forests on small-scale data sets is superior to Bimodal Distribution Removal algorithm.

Table 3. Analysis of different model's performance

Task	Unit	Original NN	NN with BDR	NN with ISR	NN with GA +BDR	NN with GA+ ISR
Classification	Accuracy	57	50	57	78.57	71.42
Regression	MSE	0.29	0.72	0.21	0.0230	0.0279

The combination of genetic algorithm optimization and BDR & ISR have significantly improved the test results than all previous models. The accuracy has increased by more than 20% and the mean square error has dropped by an order of magnitude. This shows that the combination of genetic algorithm and two removal algorithms can significantly improve the effect of neural network.

But this does not fully explain that BDR is better than ISR, because genetic algorithms train a large number of models and always choose those models that perform better, and BDR models with poor performance are not selected. Therefore, the stability of the BDR algorithm is not shown. However, this also prove that the combination of genetic algorithm and BDR can overcome the shortcomings of using BDR alone, and genetic algorithm has significantly improved BDR and ISR.

4 Discussion

Models that use neural networks alone and BDR and ISR have overfitting on classification problems, that is, the performance on the training set is acceptable, but the performance on the test data set is not good. The neural network model is good at predicting the test score whether it is the test set or the train set. The effect of the isolated forest removal algorithm is better than the other two. The performance of isolated forest removal on small-scale data sets is better than Bimodal Distribution Removal algorithm. The limitation of the Bimodal Distribution Removal algorithm is sensitive for threshold and parameter so its stability on the small-scale data set is not as good as isolated forest removal. The isolated forest removal algorithm can control the proportion of abnormal data by changing the threshold, and its performance is better. Besides, BDR only take error (loss) into consideration, while sometime neural networks have learned from the outliers and its error will not be high while ISR take original data into consideration. Therefore, it is used on the task small-scale data set ISR has more advantages.

However, when applying genetic algorithms, the shortcomings of BDR are overcome, and the performance of both BDR and ISR genetic algorithms have been significantly improved, which shows that the combination of genetic algorithms and the two is a very effective method on small-scale data sets. This combination can significantly improve the training stage for neural networks on small scale data set

The disadvantage of this study is that the data set is too small, so the results of the study have a certain randomness, that is, ISR cannot find abnormal data during training, and sometimes it needs to be repeated multiple times to get the results. There are not enough research problems, there is only one regression and one classification problem, and the classification problem is not very good. Such a result does not fully prove the conclusion. Deep learning on small-scale data sets is very challenging, and many people are involved

1. Data Enhancement: For example, if you have a picture of a cat, and the image is still a picture of the cat after rotating it, this is a good data enhancement.
2. using Cosine Loss: When converting the loss function from classification cross-entropy loss to cosine loss in the classification problem, the accuracy of the small data set is improved by 30%[8]
3. Try to use GANs to generate new data[9]
4. Modern Neural Networks Generalize on Small Data Sets[10]

In future work, you can try to increase the data set or use generative adversarial networks to expand the research and convert the regression problem into a classification problem. These jobs can be better explored.

References

1. Sörqvist P, Halin N, Hygge S. Individual differences in susceptibility to the effects of speech on reading comprehension[J]. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 2010, 24(1): 67-76.
2. Copeland L, Gedeon T. Visual Distractions Effects on Reading in Digital Environments: A Comparison of First and Second English Language Readers[C]//Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction. 2015: 506516.
3. D. Goldberg. Genetic algorithms in search[J]. optimization & machinelearning, 1989.
4. Slade P , Tamás D. Gedeon. Bimodal Distribution Removal[C]// New Trends in Neural Computation, International Workshop on Artificial Neural Networks, IWANN '93, Sitges, Spain, June 9-11, 1993, Proceedings. Springer-Verlag, 1993.

5. Liu F T, Ting K M, Zhou Z H. Isolation forest[C]//2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008: 413-422.
6. Hariri, Sahand, Matias Carrasco Kind, and Robert J. Brunner. "Extended isolation forest." arXiv preprint arXiv:1811.02141 (2018).
7. S. Luo, L. Luan, Y. Cui, X. Chai, Z. Wang and Y. Kong, "An Attribute Associated Isolation Forest Algorithm for Detecting Anomalous Electro-data," 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 3788-3792, doi: 10.23919/ChiCC.2019.8866495.
8. Barz B, Denzler J. Deep learning on small datasets without pre-training using cosine loss[C]//The IEEE Winter Conference on Applications of Computer Vision. 2020: 1371-1380.
9. Zhang X, Wang Z, Liu D, et al. Dada: Deep adversarial data augmentation for extremely low data regime classification[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 2807-2811.
10. Olson M, Wyner A, Berk R. Modern neural networks generalize on small data sets[C]//Advances in Neural Information Processing Systems. 2018: 3619-3628.