Bimodal Distribution Removal in LSTM for Distinguish between Genuine and Acted Anger

Tianhao Yu

Research School of Computer Science
Australian National University
ACT 2601 AUSTRALIA
u5869615@anu.edu.au

Abstract. In this paper, the effectiveness of the Bimodal Distribution Removal (BDR) technology is evaluated by comparing the Long Short-Term Memory Neural Network (LSTM) and the LSTM with the BDR technic on a dataset that studies distinguishing genuine from acted angry expressions. In the process of LSTM training, whether the dataset is clean or not will directly affect the training results. If there are many outliers in the dataset, it will greatly affect the training process. Here, use BDR to detect and throw out these outliers during training. Research about the efficacy and safety of BDR is limited. This experiment shows that when the model is overfitting due to the improper hyper-parameter, the BDR algorithm can achieve discard a small amount of outliers, thereby avoiding the overfitting and improving the accuracy of the model. After adjusting the hyper-parameters, the generalization of the model is not directly related to the BDR algorithm, and the final accuracy reached 94.24%.

Keywords: Bimodal Distribution Removal, Outlier Detection, Overfitting, LSTM

1 Introduction

Facial expressions are the movement of facial muscles. During social development, people learned how to control their facial expressions. Therefore, certain emotions can be forged [13]. We call those artificial and deliberate anger false angers. In this scenario, it is important to accurately distinguish between real and false anger to understand the emotional state hidden in the meaning of anger. Psychology researchers usually measure the micro-expression of displayers to discriminate genuine from fake anger [16]. In our work, we analyzed the observer's pupils diameter reaction (PDR) when watching the angry video on the monitor.

This a binary classification problem distinguishes whether anger is genuine or pretended by the characteristics of the pupils of the participants. There are many commonly used algorithms in the binary classification problem, including decision trees, random forests, and Bayesian networks [1]. In this experiment, the neural network method was used to make the classifier. First, there is a large enough dataset for training, and secondly, the neural network is robust and not easily affected by noise. Finally, this experiment uses time series data as the data set. The performance of LSTM neural networks is usually better than recurrent neural networks and hidden Markov models (HMM) [7]. In summary, in the end we used LSTM in this experiment.

The data set used in this experiment collected from 22 participants' verbal and pupillary responses when watching two anger stimuli. These two responses are used to test participants whether they can distinguish between genuine and acted anger. The accuracy rate of participants' verbal responses is only 60%. However, some researchers (Chen, 2017) used machine classifiers to learn the pupil responses of the participants, increasing the accuracy to 95% [5]. In this experiment, we tested the same problem using LSTM. Due to the characteristics of the human body, the original PDR signal will be affected by noise, such as blinking, bright, and dark of the light [2]. In order to remove noise, the original PDR signal is normalized, but there are still many outliers in the data. The pupil diameter of the experiment participants will be lost in the video frame, which may be caused by equipment and other reasons.

Outliers refer to the fact that one or more values in the data differ greatly from other values. Outliers will greatly affect the training of LSTM. When using LSTM for evaluation, outliers can cause large errors. When a gradient-based learning method is used, this large error will result in a large weight update, which pushes the network weights to outliers, thereby adversely affecting the learning of the LSTM [17]. When there are more outliers (> 5%), the model will be greatly distorted. In this case, the model is prone to be unable to converge or overfit [4]. To solve this problem, researchers have proposed many methods, including Isolated Forest, Least Median Squares (LMS), and Least Trimmed Squares (LTS) etc. Most of these methods perform not very well on real-word data. *Bimodal Distribution Removal* (BDR) research shows that with the training epoch increases, the frequency of prediction errors in the training set appear a bimodal error distribution. As shown in Figure 1, the peak with high error contains the outlier. Bimodal Distribution Removal (BDR) has been used for removing the outliers from the training dataset during the training

process. This method also provides a natural stopping to halt the training, thus greatly avoiding overfitting [15]. To date, BDR has received scant attention in the research literature. This paper studies whether the BDR algorithm can play a positive role in the training process and results when the data set has more noise.

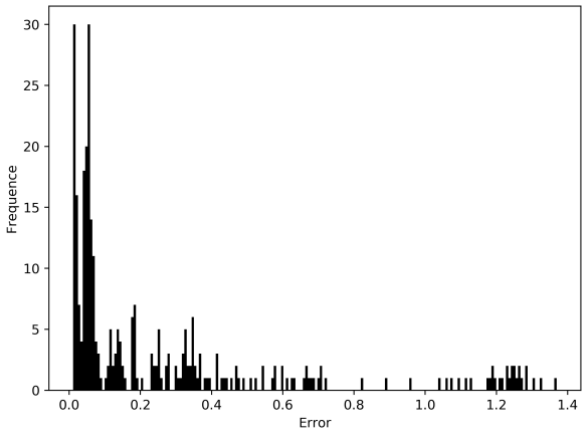

Fig. 1. Error distribution at epoch 20 in histogram. The error is calculated by using Cross entropy loss function.

2 Method

2.1 Aim

This article will examine the effectiveness of LSTM in classifying genuine and acted anger under the input pupil diameter of participants. In particular, I will test the effectiveness of BDR to detect outlier patterns and the ability of the algorithm to filter out noise features to determine whether it contributes to the performance of LSTM.

2.2 Data Acquisition and Signal Processing

Data pre-processing refers to some processing of data before the main processing. In order to improve the quality of the LSTM, there are mainly three kinds of pre-processing have been done. The first step in this process was to eliminate more than the data in the data set, only retain the features and labels, and ensure that the data type can be used for training. For example, conversion labels. In the dataset, the label is presented as “Tx” and “Fx” which cannot be trained by LSTM. Thus, these two labels are assigned as 1 and 0. However, there are still many missing values in the data at this time. We chose to discard the data that experimental data were lost by more than 50% due to the light and the pupils of the participants themselves. There is also some experimental data loss in some frames, which is caused by factors such as experimental equipment. For this situation, we use linear method to interpolate these data. The pupil data is divided into left and right pupils, and here we average the two sets of data. Finally, in order to eliminate the dimensional impact between the indicators, solve the problem of comparability between data indicators. We also did Min-Max normalization to each participant separately to reduce their interval difference and speed up the convergence of LSTM. The length of the video is inconsistent, which makes it impossible to use the mini batch method for training. Therefore, we padded zeros at the end of each sequence to keep their length consistent. Then, the data set is divided into a training set and a test set according to a ratio of 80%, which are used for training and testing LSTM, respectively. The formula used for normalization is showing below. All the above operations are done by libraries in Python, including NumPy, Pandas, and PyTorch.

$$\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

2.3 Feature Selection

Due to the very long interval and delay in the time series of our data, it is more appropriate to use LSTM [9]. If the Recurrent Neural Network is used, the gradient disappears easily. In our experiment, there are two Long-Short Term Memory neural network (LSTM) models trained. They trained by the same training set and same hyper-parameters. The

only difference between them is whether using *Bimodal Distribution Removal* (BDR) for outlier pattern removal or not. In the experiment, we would call the LSTM without BDR as control group, and the LSTM with BDR as experiment group.

To have a good classifier, we designed a 2-layer LSTM network consisting of 31 input neurons and 2 output neurons, because and this is a binary classification problem. Here, if the number of input size is too small, the problem of vanishing gradient is easy to occur, resulting in the model not converging. Tthe chosen input size of LSTM is 31. In this case, the LSTM will unroll six times, because the sequence length of our data is 186. As this is a binary classification problem, there is no need to use softmax as the activation function in the last hidden layer [3]. The cross-entropy loss is very suitable for predicting classification problems, so here we only use the cross-entropy for loss function [10]. Here we use Adam as our optimizer, because Adam is suitable for the problem of gradient sparseness or data with large noise, and can naturally accomplish the step annealing process [9]. Through continuous testing, the hyper-parameters of the neural network are determined by the rule of thumb, including the number of hidden neurons, the number of epochs, and batch size.

2.4 Bimodal Distribution Removal Implementation

In 1993, Gedeon and Slade proposed algorithm of *Bimodal Distribution Removal* to remove the outlier in the training set during the training. The algorithm needs to set three constants as thresholds. One of the constants, α is used to distinguish whether the pattern is an outlier or not, and the other constant, β is used to decide when to halt the training. The formula use to determine the patterns (outliers) need to be remove is showing Formula 2, where δ_{ss} and σ_{ss} is the mean and standard deviation of the loss in certain epoch. The pattern will be removed from the training set, when error of this pattern is greater than the mean plus standard deviation times α. In bimodal or skewed distribution, the pattern is more than one standard deviation away is defined as outlier [14], therefore it is reasonable to set the constant α between 0 and 1. The variance of loss for whole training set, v_{ts} is calculated each epoch to determine whether the training need to halt. The training will halt when the variance of loss for whole training set, v_{ts} less than constant β (0.001) [15]. The other constant is a threshold used for detecting whether the probability distribution of the error forms a bimodal distribution during the training.

As shown in Figure 1, the error distribution at epoch 0 exhibits a bimodal distribution. For our experiment, implementation of BDR technic need to initiate after several epochs, usually after 30 to 40 epochs. As mentioned above, the outlier in the training dataset can be detected by the variance of the training loss.

$$error \geq \overline{\delta}_{ss} + \alpha\sigma_{ss} \qquad where\ 0 \leq \alpha \leq 1 \tag{2}$$

2.4 Experiment Setup

The training process is implemented by PyTorch 1.4.0. The program is performed with an AMD Ryzen™ 7 1700 with 3 GHz, 16.00 GB of RAM, Operating System 64-bit computer using Python 3.7.5. In order to test the performance of the classification, the dataset is divided by Pandas and used for training, validation and testing respectively. After continuous testing and adjustment of hyper-parameters, the trained model performs the best, when the number of neurons in the input layer is 31, the number of neurons in the hidden layer is 32, the number of epochs is 300, and the learning rate is 0.1. The same hyper-parameters applied for both control and experiment group. The batch size is set to be 64. Theoretically, the effect of full batch on small dataset should be better. However, the model is easy to overfit on the outlier pattern, due to the problem of noise [6]. As shown in Table 1, these experiments confirmed that the model performs the best at batch size is 128 under the situation of all other hyper-parameters are the same. This Table also reflects from the side that there are a large number of outliers in datasets. Other hyper-parameters are adjusted in the same way. To observe the effect of BDR technic on the modal training. Both the experimental group and the control group trained 40 different models and compared the accuracy on the test set.

Table 1. Comparing the Accuracy on test dataset with different Batch size

Batch size	Accuracy (%)
32	89.46 (±2.40)
64	**92.82 (±2.64)**
128	91.74 (±2.32)
Full batch	92.12 (±2.13)

3 Results and Discussion

Through the control variable method, we compared the results of the control group and the experimental group. Experimental results were used to determine whether the BDR improves the generalization of the model and reduces the training time by removing the outlier pattern. As mentioned in the previous section, we use the previously defined hyper-parameters to train with LSTM and LSTM with BDR. The data of the training set for each experiment are randomly selected from the dataset. After several experiments, the LSTM model is compared with the testing accuracy with and without the BDR technic. The BDR technic does greatly reduce the training time of the model. After repeated experiments, it was found that the BDR technic did greatly reduce the training time of the model. In addition, BDR can increase the accuracy of the model by throwing out amount of outliers and early stopping methods, but this needs to be achieved by adjusting some of the hyper-parameters in the BDR algorithm.

3.1 Effectiveness of LSTM Feature Selection

The difference between the test results of each experiment is relatively large, this is due to the random initialization of the weight parameters and the difference between training set in each training and other reasons. To ensure the validity of the test results, the cross-validation method has been used here. We took 10 experiments to take the mean and standard deviation and discarded the outlier. After 10 experiments on the same set of hyper-parameters mentioned above, the accuracy of LSTM on the test set reached 94.14%. Figure 2 shows the accuracy of the test set and the training set in each epoch. It can be seen from the figure that the model has completed convergence at about 120 epochs, but the accuracy on the test set has not continued to improve. The training error also remains the same, indicating that there is an overfitting situation. Overfitting is an analysis result that corresponds too accurately to a particular data set, so it may not be possible to fit other data or reliably predict future observations [8]. Due to the loss of some data in the training data set, although these data have been filled by linear interpolation, it still causes a large amount of noise in the data set. Therefore, the model overly remembers the noise characteristics but ignores the relationship between the actual input and output [12]. There are only 310 pieces of data in the training set. Insufficient training data will cause the model to learn completely irrelevant information (such as noise) from the training data.

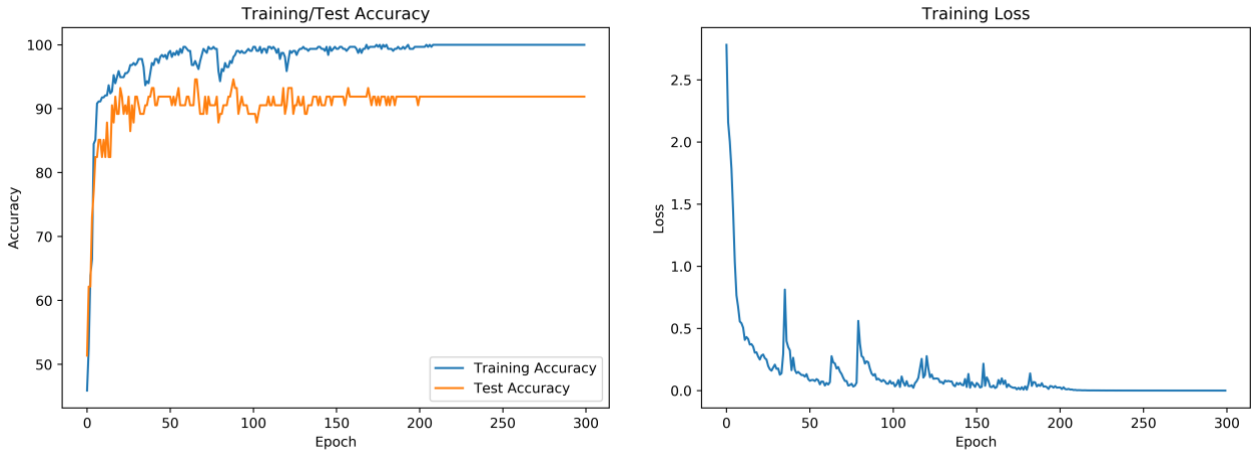

Fig. 2. The left plot is Training and Test Accuracy for LSTM and right plot is Training loss of LSTM

3.2 Effectiveness of LSTM with BDR Feature Selection

In Figure 3, we set the α to 0.8 and to 0.001. Under this parameter setting, the average performance of the model can reach the best. After using BDR technic on LSTM, the average accuracy on test set reaches 94.24 (±2.74) %, and the training accuracy is 96.92%. After implement BDR on LSTM, the training usually halts around 55 epochs. Through experiments, I can find that BDR can indeed reduce training time by providing a natural stopping. Figure 3 indicates that after joining the BDR technic, LSTM did finish training early at the highest point of training accuracy and the lowest point of training error. With the increase of epoch, the model will improve the adaptability to the training data but at the cost of increasing the generalization error. In theory, early stopping will indeed increase the generalization ability of the model [18]. Figure 4 confirms that the probability of errors during training will indeed exhibit a bimodal distribution. The threshold setting is very important, when using BDR to detect this bimodal distribution during the training. The threshold of variance does not have much effect on the accuracy of the model, and it does not directly prove that ending the training early can improve the accuracy of the model.

Table 2. Experimental data is used to find the most suitable constant

α (sigma distance)	β (threshold of variance)	Accuracy
0.2	0.001	87.21 (±3.21)
0.4	0.001	88.34 (±3.35)
0.6	0.001	90.28 (±2.52)
0.8	0.001	93.24 (±2.74)
0.8	0.003	93.01 (±2.40)
0.8	0.005	93.21 (±2.37)
0.8	**0.007**	**94.24 (±2.74)**
0.8	0.008	93.33 (±2.41)
0.8	0.01	93.16 (±1.40)

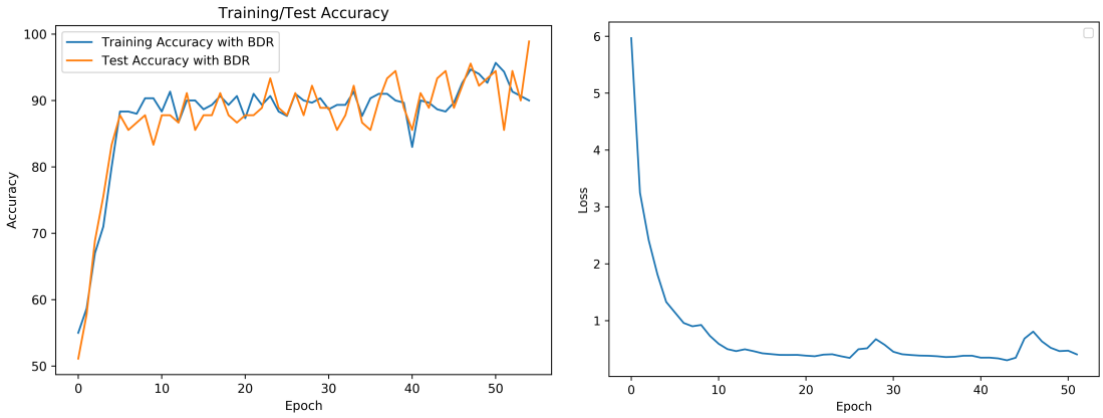

Fig. 3. The left plot is Training and Test Accuracy for LSTM with BDR and right plot is Training loss of LSTM with BDR at α =0.8 and β = 0.007

In Figure 3, we set the α to 0.8 and to 0.007. Under this parameter setting, the average performance of the model can reach the best. After using BDR technic on LSTM, the average accuracy on the test set reaches 94.24 (±2.74) %, and the training accuracy is 96.92%. After implementing BDR on LSTM, the training usually halts around 100 epochs. Through experiments, I can find that BDR can indeed reduce training time by providing natural stopping. Figure 3 indicates that after joining the BDR technic, LSTM did finish training early at the highest point of training accuracy and the lowest point of training error. With the increase of epoch, the model will improve the adaptability to the training data but at the cost of increasing the generalization error. In theory, early stopping will indeed increase the generalization ability of the model [18]. Figure 4 confirms that the probability of errors during training will indeed exhibit a bimodal distribution. The threshold setting is very important when using BDR to detect this bimodal distribution during the training.

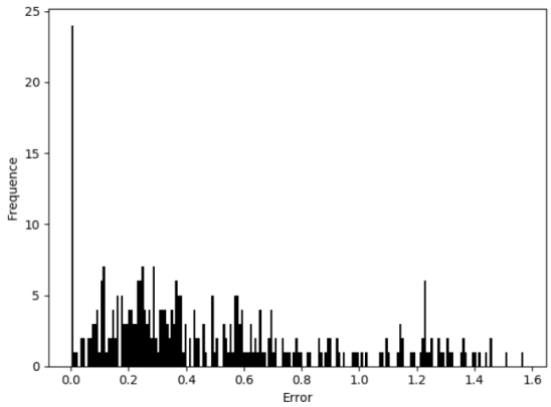

Fig. 4. Error distribution formed during the training at epoch 22

Through the comparison in Figure 5, the BDR algorithm accelerates the convergence rate of the model by discarding some outlier patterns, under the same hyper-parameter settings. The accuracy rate on the test set is also increased by 1.5%, which reflects from the side that BDR can indeed improve the generalization ability. However, as shown in Table

2 shows that the prediction accuracy of the model is not increased with the more patterns discarded. A large part of the BDR removal mode is effective input features, and removing these features will naturally lead to a decline in prediction. The BDR technic is very sensitive to parameter setting, only when the parameter setting is reasonable can outliers be better removed. This requires a lot of tuning to find the parameters suitable for the model and data set. In this experiment, in order to get a better prediction on the test set, I modified a lot of preset value, that Tom and Slade mentioned the preset value in the paper [15].

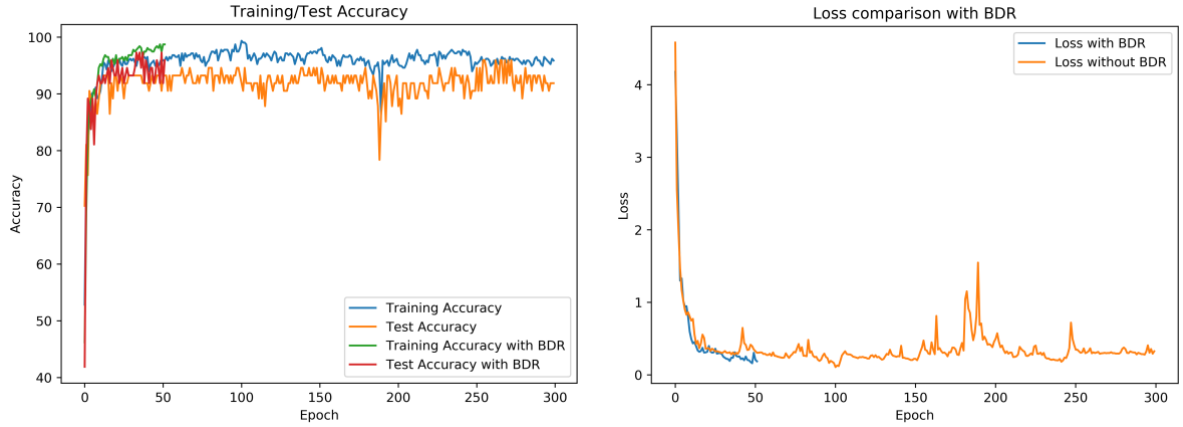

Fig. 5. Comparison of Training and Testing Accuracy and Training Loss between LSTM and LSTM with BDR.

Although, the training halted at the peak of train accuracy and bottom of the training loss. Figure 5 shows that after the BDR discards some outliers, the accuracy of the model on the test set has been significantly improved. It shows that the removal of outliers by BDR does improve the generalization performance of the model to a certain extent. It can also be seen that the LSTM model is difficult to fit on the anger dataset of this experiment. This aspect is caused by the setting of hyper-parameters. Methods such as increasing the number of neurons in the hidden layer can indeed make the model fit last, but also increase the possibility of overfitting. On the other hand, the model does not fit well because of too many outliers in the data set. Therefore, after discarding a small amount of outlier through BDR, the accuracy of the model can be improved. Whether the variance of the training error can be used for the criteria of early stopping remains to be studied. As Figure 5 suggests that there is still a small amount of "slow coach" that can be learned from the training set.

Ke also did some research on BDR [11]. Most of our conclusions are the same, but the results are not the same. This may be caused by difference between the dataset. The data set used by Ke may be relatively "clean" (have fewer outliers). In this way, he will have a great possibility to remove the effective features, when using BDR technic. Therefore, the accuracy of model prediction has decreased significantly after using BDR.

4 Conclusion and Future work

From the results, after carefully setting the threshold, BDR can indeed slightly improve the generalization ability of the model and reduce the training time for the model by comparing it to the control group. The early stopping method of BDR can reduce training time. However, whether the early stopping method has improved the accuracy of the model still needs further proof. Although the study has successfully demonstrated the effectiveness of the BDR, it has certain limitations in terms of setting the parameter for BDR, including the sigma distance (). The BDR technic is too sensitive to parameter setting, only when the parameter setting is reasonable can outliers be better removed. This leads to detect outliers very difficult during training, on the other hand, data pre-processing, dropout, and pruning are more effective in preventing overfitting. After using other methods, if the model is still overfitting, BDR can be tried as an alternative.

This paper discusses the effectiveness of the bimodal distribution removal method in improving the generalization performance of LSTM classification problems. The dataset used in this paper is the time series. In the future, I will try to use Convolutional Neural Networks (CNN) to improve the accuracy of the model prediction on image classification problems. Then add BDR technic based on CNN and observe whether it improves the generalization of the model. Also, I will try to find new values to define the natural stopping point, and to get a more accurate model from learning "slow coach".

References

[1] Abend, K., Harley, T., Kanal, L.: Classification of binary random patterns. IEEE Transactions on Information Theory. 11, 538-544 (1965).
[2] Becht, J.: Bright Pupils and Dull Pupils. Journal of Education. 79, 395-396 (1914).
[3] Bridle, J.: Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. Neurocomputing. 227-236 (1990).
[4] Bullen, R., Cornford, D., Nabney, I.: Outlier detection in scatterometer data: neural network approaches. Neural Networks. 16, 419-426 (2003).
[5] Chen, L., Gedeon, T., Hossain, M. Z., & Caldwell, S.: "Are you really angry?: detecting emotion veracity as a proposed tool for interaction." In Proceedings of the 29th Australian Conference on Computer-Human Interaction (pp. 412-416). ACM. (2017)
[6] DeWeese, M.: Optimization principles for the neural code. Network: Computation in Neural Systems. 7, 325-331 (1996).
[7] Gers, F.: Learning to forget: continual prediction with LSTM. 9th International Conference on Artificial Neural Networks: ICANN '99. (1999).
[8] Hawkins, D.: The Problem of Overfitting. Journal of Chemical Information and Computer Sciences. 44, 1-12 (2004).
[9] Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation. 9, 1735-1780 (1997).
[10] Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2014).
[11] Quan, K.: Bimodal Distribution Removal and Genetic Algorithm in Neural Network for Breast Cancer Diagnosis. (2020).
[12] Pham, H., Triantaphyllou, E.: The Impact of Overfitting and Overgeneralization on the Classification Accuracy in Data Mining. Soft Computing for Knowledge Discovery and Data Mining. 391-431 (2008).
[13] Weker, M.: Smile and Lie? Why We Are Able to Distinguish False Smiles from Genuine Ones. Issues in Science and Theology: Do Emotions Shape the World?. 59-71 (2016).
[14] Šaltenis, V.: Outlier Detection Based on the Distribution of Distances between Data Points. Informatica. 15, 399-410 (2004).
[15] Slade, P., Gedeon, T.: Bimodal distribution removal. New Trends in Neural Computation. 249-254 (1993).
[16] Strongman, K.: The psychology of emotion. Wiley, Chichester (1998).
[17] Van der Stockt, S.: A generic neural network framework using design patterns. (2005).
[18] Zhang, T., Yu, B.: Boosting with early stopping: Convergence and consistency. The Annals of Statistics. 33, 1538-1579 (2005).