Bimodal Distribution Removal Experiment with Semeion dataset

You Wu,

Research School of Computer Science, Australian National University
u6023244@anu.edu.au

Abstract. We try to explore how the Bimodal Distribution Removal method [1] affects the results of a convolutional neural network. The dataset we picked from UCI is the Semeion Handwritten Digits dataset [2]. We selected a CNN model with 2 convolutional layers, 2 max pooling layers and 1 output player for the classification task. After applying the BDR method, the accuracy of classification on test dataset locates in a range of 93% ~ 96%, which is better than the results in the Automatic Representation and Classifier Optimization for Image-based Object Recognition [3]. We also discovered that the Semeion Handwritten Digits dataset [2] contains few outliers, which leads to that the BDR method affects little on the result.

Keywords: Bimodal Distribution Removal, Semeion Handwritten Digits, Convolutional Neural Network, Test set method

1 Introduction

The main purpose of the experiment is to evaluate the influence of the Bimodal Distribution Removal [1] on the results of a convolutional neural network which performs a classification task on a typical grid-like dataset. The dataset picked is the Semeion Handwritten Digits dataset [2] which has several advantages for the experiment and can bring an efficient training procedure. The dataset is typical for classification task as it consists of handwritten digits images. We firstly tried to find a suitable convolutional neural network model by the test set method for the classification task and then tried to apply the Bimodal Distribution Removal [1] on the model and evaluate the result. The BDR method is determined to have few impacts on the accuracy of results. But it reduces the number of epochs needed in for training and prevents overfitting.

1.1 Dataset and problem model

The dataset picked from UCI Machine Learning Repository is called the Semeion Handwritten Digits dataset [2]. The dataset was created by Tactile Sri, Brescia, Italy and donated in 1994 to Semeion Research Center of Sciences of Communication, for machine learning research. It consists of 1593 scanned image of handwritten digits from around 80 persons, stretched in a rectangular box 16x16 in a gray scale of 256 values. Each pixel of each image was scaled into a boolean (1/0) value using a fixed threshold. Each person wrote on a paper all the digits from 0 to 9, twice. The commitment was to write the digit the first time in the normal way (trying to write each digit accurately) and the second time in a fast way (with no accuracy).

There are 2 advantages of the dataset for this experiment. The first one is that its features are 256 columns of gray-scale values, which actually describe a 16*16 gray-scale image. It means the input will be grid-like, which is suitable for CNN. The second advantage is that the dataset is not too large like the MNIST dataset. It contains 1593 instances and the size of the images is 16*16, which means the procedure of training will not be too long.

1.2 Pre-process

The Semeion Handwritten Digits dataset [2] consists of 266 columns. The first 256 columns are the scaled gray-scale values of the scanned handwritten digits images. We take the first 256 columns of each instance out as a vector and reshape them as 16*16 matrices.

The last 10 columns indicate which number is in the images. For an instance, the value in the corresponding column of the number in the image in the last 10 columns will be 1 while the values in the rest 9 columns will be 0. For example, if there is a instance in which the number in the image is 0, the value in the 257^{th} column will be 1 and the value in the 258^{th} column to 266^{th} column will be 0. However, the labels we need for build a CNN performing a classification task are a single column contains numbers in a range of 0 ~ 9. Thus, we traverse the last 10 columns of each instance and determine which class they belong. Then, we replace the last 10 columns with a single column which contain a number in range of 0 ~ 9 indicating which class the instances belong.

Eventually, the raw dataset is processed and becomes a two-column matrix. The first column contains a 16*16 matrix for each instance, which represents the scaled gray-scale image of the scanned handwritten digit. The second column contains an integer in the range of 0 ~ 9 for each instance, which indicates the number in the image. Then, the first column is used as input and the second column is used as target. In the experiment program, we split the dataset into training set (80%) and testing set (20%) randomly.

1.3 Problem to solve

The 'bias and variance' dilemma is a famous dilemma in the neural network field. One approach to relieve the dilemma is to remove the outliers from the dataset to reduce the initial variance in the training set and thus improve the variance/bias tradeoff. On the other hand, it can reduce the size of training set during training procedure. With a smaller training set and those outliers which are hard for the neural network to learn about, efficiency of training neural network will get a great improvement. The outlier detection method researched in this experiment is Bimodal Distribution Removal [1]. By applying the method on our neural network model and compare the results, we explore the effect of the method on the specified dataset and method. The results will be compared in several aspects such as the trend of loss on training set and test set during training, the accuracy of the prediction on test set and the loss frequency distribution of the instances in training set.

2 Method

The main target of the experiment is to explore the influence of the Bimodal Distribution Removal [1] on the results of our convolutional neural network when it performs a classification task on the Semeion Handwritten Digits dataset [2]. Before applying the BDR method, we need to pick a suitable network model for the classification task. With consideration of efficiency, we decide to choose model by test set method.

The model picked is CNN with two convolutional layers, two max pooling layer and one output layer, which depends on the comparison between results of several different CNN models. With the suitable network model, the BDR method is applied to detect the outlier in the dataset and improve the efficiency of training.

2.1 Model determination

As mentioned, input for the task is consists of 16*16 images. It means we can use at most 2 combinations of convolutional layer and max pooling layer, otherwise the feature map will be too small and meaningless. The task is modeled as a classification task and the instances in the dataset belong to 10 classed as they are images of handwritten digits in a range of 0 ~ 9, which means 10 output units are needed.

The emphasis of the model is the structure of hidden layers. We tried several different schemes of hidden layers and compare their results on the dataset. The aspects of results we compared are the trend of loss on training set and test set, the minimal loss on test set and the number of epochs needed before overfitting occurs.

We implemented a CNN with 1 convolutional layer, 1 max pooling layer, 1 linear hidden layer and 1 output layer at the beginning. Based on experience, the learning rate is set to 0.001. The result of this model is typical. It seems that there is no special issue in the dataset.

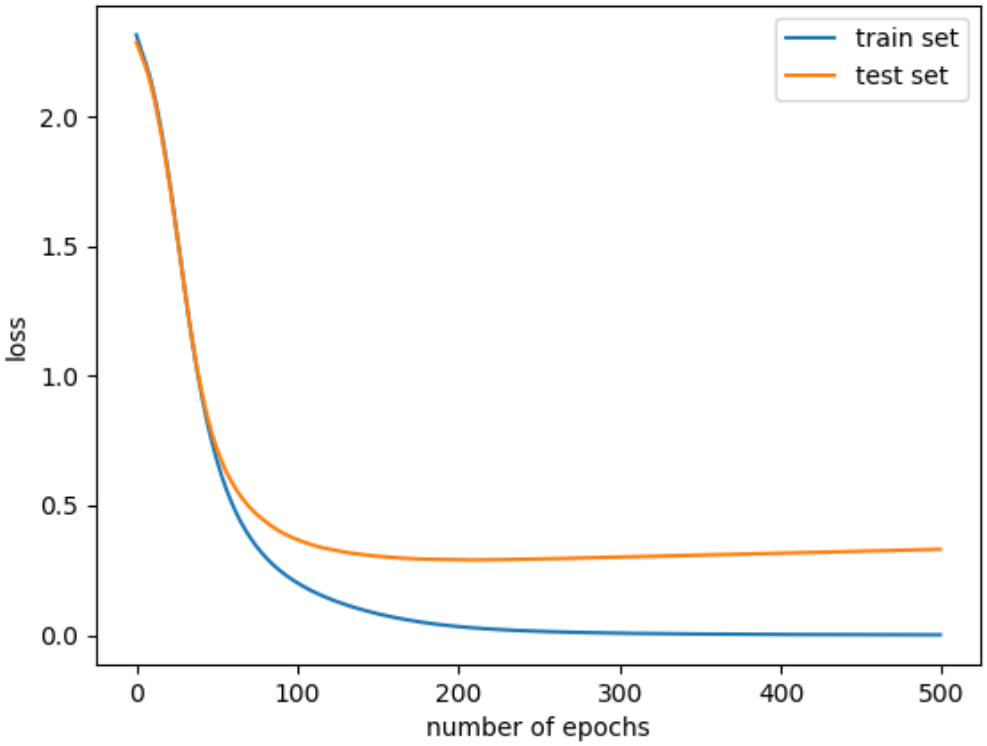

Fig. 1. The trend of loss on training set and test set of the first model. There are 1 convolutional layer with 10 output channels, 1 max pooling layer, 1 linear hidden layer with 50 neurons and 1 output layer.

According to the above figure, the first model needs approximately 200 epochs before getting the overfitting point. The value of minimal loss on test set is approximately 0.3. It seems that the first model is a good model, but we still need to compare it to other models to find a better model.

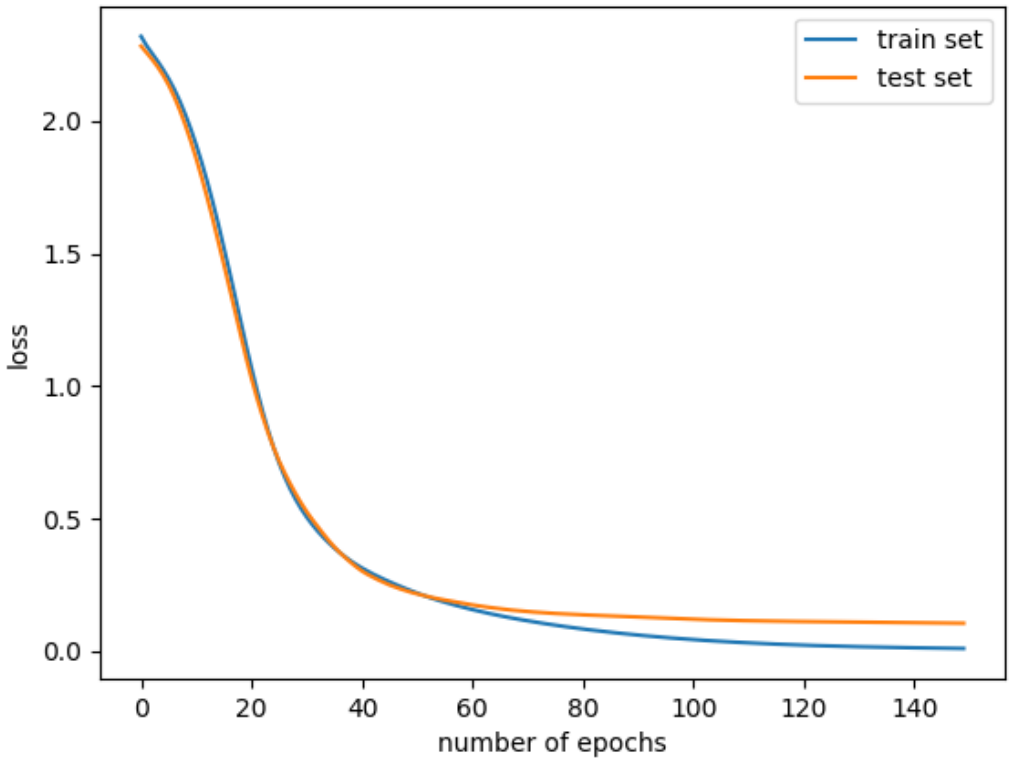

Fig. 2. The trend of loss on training set and test set of the second model. There are 1 convolutional layer with 15 output channels, 1 max pooling layer, 1 convolutional layer with 30 output channels, 1 max pooling layer and 1 output layer.

Compared to the first model evaluated before, it needs about 100 epochs before overfitting occurs, which is much quicker than the first model. On the other hand, the minimal loss on test set for this model is approximately 0.1, which is also much better than the first one. Since this network has a faster convergence than the first one and perform better in the minimal loss on test set aspect, it is a better model than the three-layer network above.

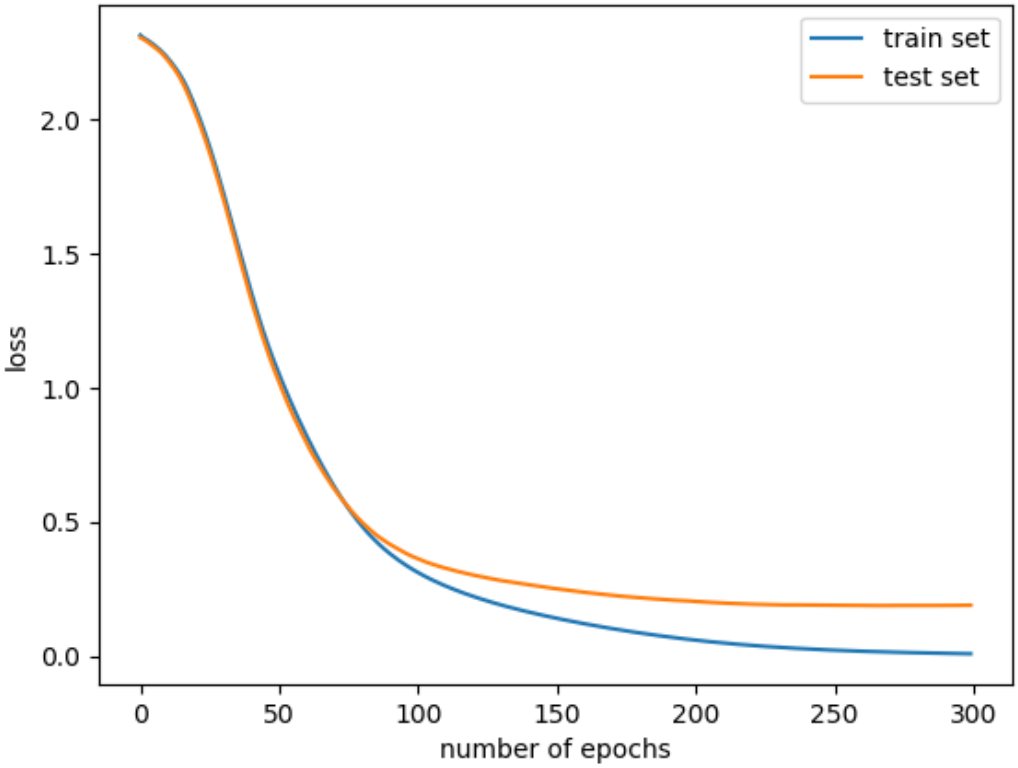

Fig. 3. The trend of loss on training set and test set of the third model. There are 1 convolutional layer with 10 output channels, 1 max pooling layer, 1 linear hidden layer with 50 neurons, 1 linear hidden layer with 20 neurons and 1 output layer.

The performance of the third model is also explored with the learning rate being 0.001. An extra linear hidden layer does not improve much on the results. It takes approximately 200 epochs to getting the minimal loss on the test set, which means it needs more training time with the first model as there are more neurons. The minimal loss on the test set for this model is about 0.1, which is a little better than the first model.

We also try to add dropout to the models. It brings a lot of oscillations to the trend of loss. But it does not improve the models much.

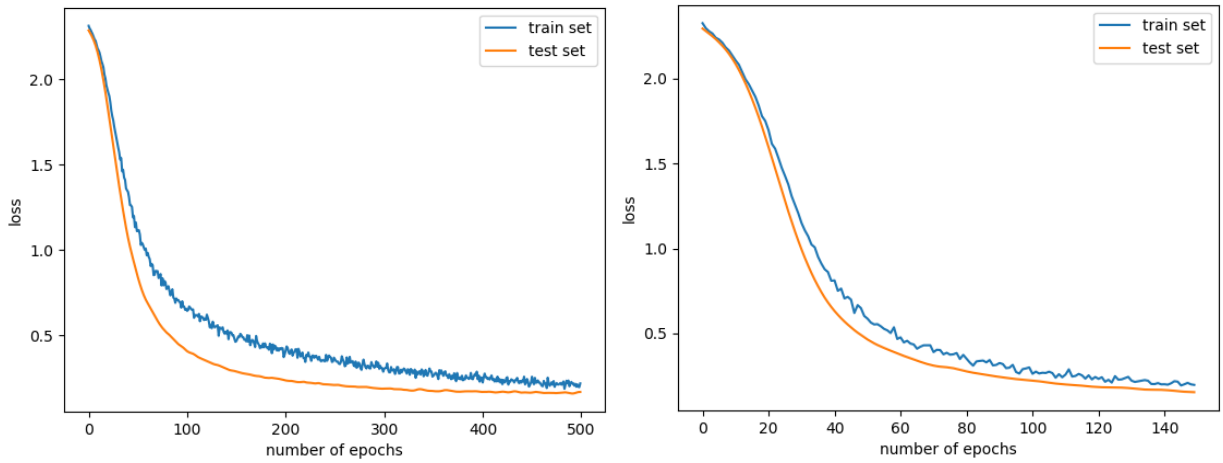

Fig. 4. The trend of loss on training set and test set of the first model and the second model with dropout.

From the above figure, we can not find improvement on convergence speed or minimal loss in the results of the first model and the second model with dropout.

Generally, the second model without dropout is selected as the network model in the following research. On the basis of above comparison, it has the best training efficiency and perform best in the accuracy aspect.

2.2 Bimodal Distribution Removal

The Bimodal Distribution Removal [1] is a method for outlier detection, which remove patterns in an intentionally conservative way and has a halting criterion preventing overfitting. In most datasets, the error frequency distribution of instances will be scattered with large variance at the beginning of training procedure. However, the error for a majority of the training set is reduced pretty quickly as those patterns are learned by the network. In the error frequency distribution diagram, there will be a peak in the small error area, which contains the well learned patterns. Those patterns which are not in the low error peak are likely to be the outliers. But they should not be removed too quickly as those patterns with midrange errors could eventually be learned [1].

According to [1], we determine the patterns which are removed by several steps. Firstly, the mean error of all patterns $\overline{\delta}_{ts}$ is calculated. As the low error peak contains the majority of all patterns, the value of $\overline{\delta}_{ts}$ will be quite small but larger than the error for most patterns in the peak. Those patterns that are likely to be outliers will have a error larger than $\overline{\delta}_{ts}$, they are taken from the training set and make up a subset.

This subset contains the patterns which are not in the low error peak and not learned well by the network. It contains outliers and those patterns with midrange error, which is called slow coaches and can be learned by the network eventually. As the proportion of outliers in this subset is higher, its mean error $\overline{\delta}_{ss}$ will be large. Besides, standard deviation of the subset σ_{ss} is calculated. We will decide which patterns to permanently remove from the training set with these two statistics. Those instances with

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

are permanently removed. This inequality shows that the BDR method removes patterns in intentionally conservative as mentioned, most of patterns in the subset taken from training set will not be removed. The above steps are repeated every several epochs (typically 50) during the procedure of training, which gives the network a period to learn the features of the modified training set.

Another advantage of the BDR method is that it has a halting criterion. Outlier removal should not be continued indefinitely as all the patterns in the training set will be removed eventually in that way. During the removal procedure, the training set will become smaller and smaller, which means the possibility of overfitting will increase. The BDR method uses v_{ts} as its halting criterion. When v_{ts} is smaller than a constant (typically 0.01) training is halted [1]. It is because that the patterns in the high error peak are removed, which means that the proportion of the patterns in low error peak increases and the value of $\overline{\delta}_{ts}$ and v_{ts} will be quite small.

Generally, the BDR method has 4 advantages in comparison with previous outlier detection methods:

- pattern removal does not start until the network itself has identified the outliers,
- the number of patterns removed is not hard wired, but instead is data driven,
- patterns are removed slowly, to give the network ample time to extract information from them, and
- a halting criterion naturally evolves preventing overfitting – results in significantly faster training time [1].

3 Results and Discussion

As mentioned, the BDR method uses the variance of losses for all patterns in training set v_{ts} as the criterion of starting outlier detection and halting training. Before applying the BDR method on the dataset and network, we firstly observe

the trend of v_{ts} during training procedure. As the number of epochs is small in our training procedure, we decide to calculate variance every 10 epochs rather than every 50 epochs.

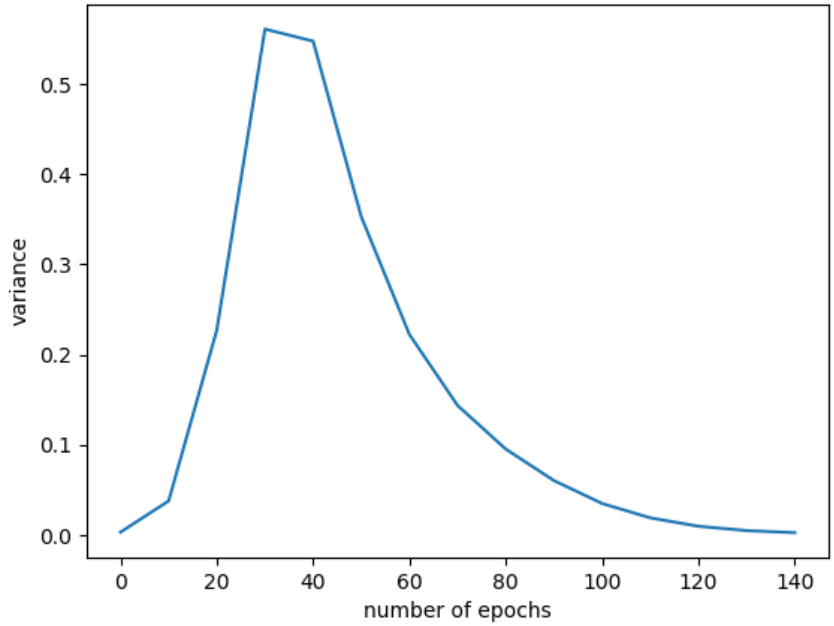

Fig. 5. The trend of error variance during training procedure.

The result is pretty different from expectation. Unexpectedly, the value of v_{ts} is quite small at the beginning of training, which almost achieves the halting criterion (0.01). In the early term of training, v_{ts} increase rapidly. After achieving its maximal value, v_{ts} is reduced very slowly.
To determine the reason for the unusual case, we decide to investigate behavior of the dataset.
After the investigation, effect of the BDR method and differences of results between our method and previous methods will be observed.

3.1 Behavior of dataset

As unexpected case occurred during model determination, the abnormal trend of v_{ts} is considered as a dataset issue tentatively. Comparison of our error frequency distribution trend with the trend in [1] will show what is the difference between behavior of the two datasets.

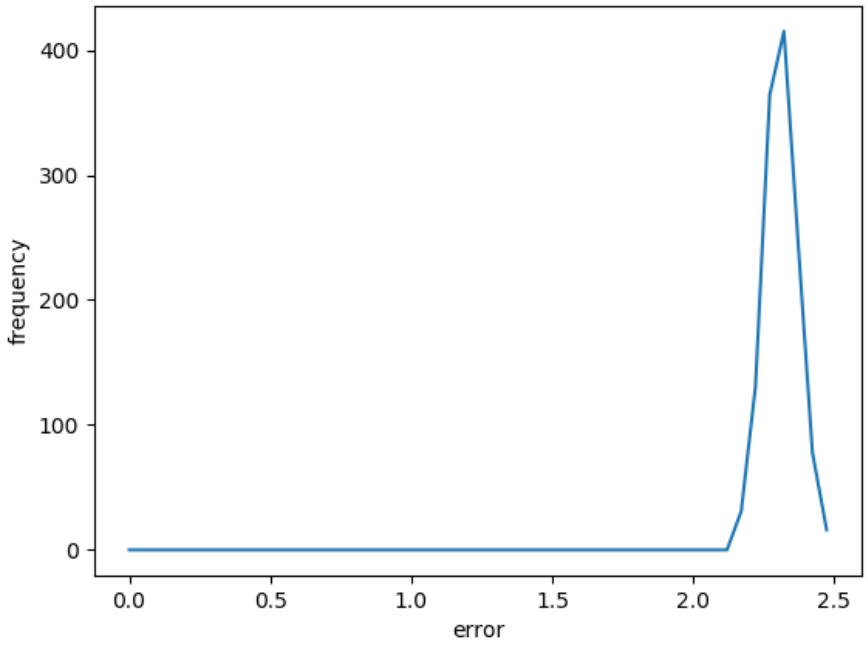

Fig. 6. The error frequency distribution of patterns in train set at the first epoch of training procedure.

At the beginning of training procedure, the error frequency distribution of our dataset is not scattered like the case in [1]. The figure shows that all the patterns have similar high error, which leads to the case that v_{ts} is quite small at the beginning of training unusually. However, the mean error for all patterns in training set $\overline{\delta}_{ts}$ is quite large, which is different from the situation after a number of epochs.

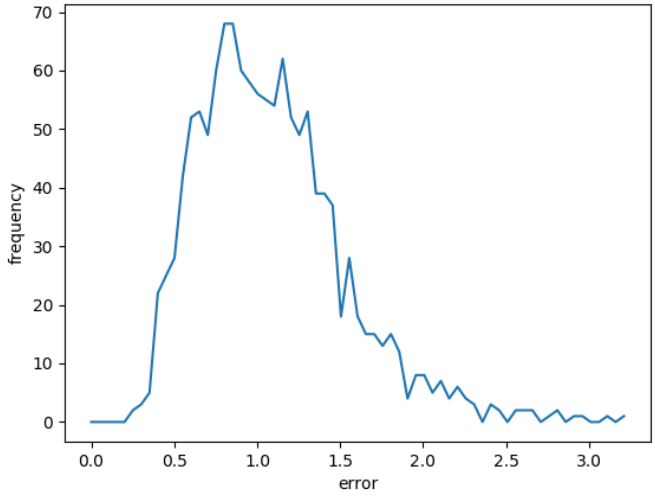

Fig. 7. The error frequency distribution of patterns in train set at the 20th epoch of training procedure.

After 20 epochs training, the error frequency distribution begins to change. A number of instances in the dataset get smaller error, which means they are learned by the network gradually. A minority of instances have high error, which is similar with [1]. In this case, the variance of error is quite large. Since there are approximately 1200 instances in the training set totally while the largest frequency in the above figure is 70. The error values distribute in a scattered way.

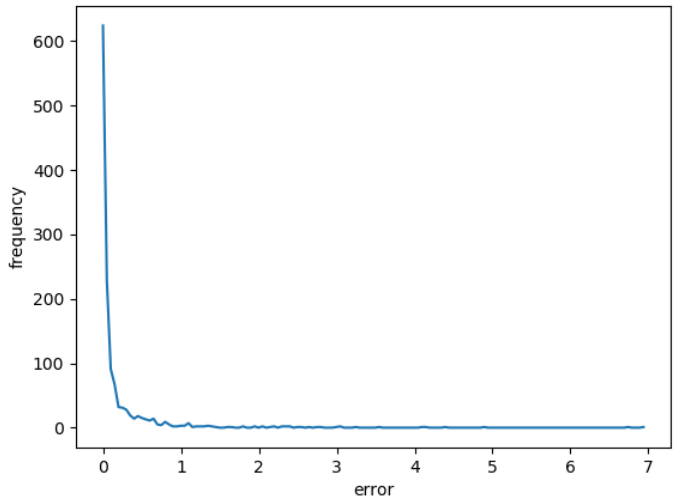

Fig. 8. The error frequency distribution of patterns in train set at the 50th epoch of training procedure.

The network begins to learn about most instances in the dataset. It looks like that all patterns are in the extremely low error peak in the above figure. But what is noticeable is that the peak contains about 600 instances, which is only half of the training set. It means there is still a number of patterns in the high error area, which is also shown by the maximal value of the X axis. It is actually a 'bimodal distribution'.

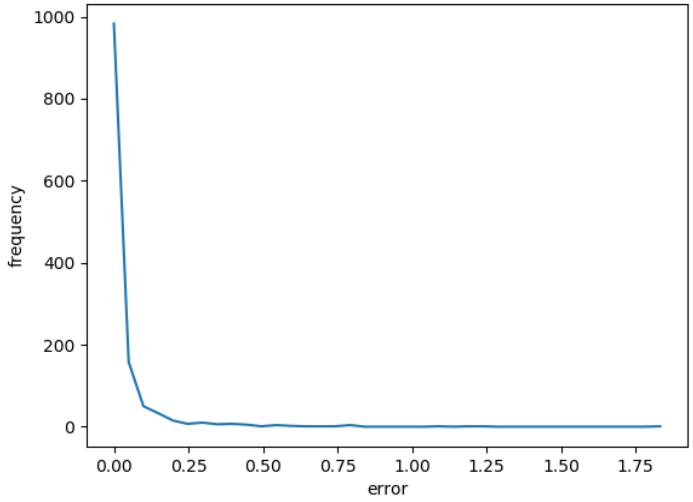

Fig. 9. The error frequency distribution of patterns in train set at the 120th epoch of training procedure.

Our distribution is much more similar to the distribution in [1]. Approximately 1000 instances are contained in the peak in the low error area. The value of v_{ts} is smaller than 0.1 in this period, which is the criterion of starting outlier detection and removal in [1], it is a reasonable timing to start removal.

On the basis of above observation, the abnormal trend of v_{ts} during the training procedure is considered as an issue caused by the property of the dataset. In practice, we decide to combine $\overline{\delta}_{ts}$ and v_{ts} as the criterion of starting outlier detection and removal, which is also used as the criterion of halting training.

3.2 Results after BDR

After observation of behavior of the dataset, we apply the BDR method [1] in the program.

It generally contains 3 steps:

- if variance of instances loss is smaller than 0.1 and mean loss is smaller than 0.3, take the instances with loss larger than mean loss from the training set,
- remove those instances satisfying $error \geq \overline{\delta}_{ss} + \alpha\sigma_{ss}$, the α in our program is 1.0,
- if variance of instances loss is smaller than 0.01 and epoch loss is smaller than 0.2, halt training.

The BDR method has few impacts on the training procedure of our network. Outlier removal happens for few times during the training procedure. But the halting training criterion is usually achieved before 150 epochs training are finished, which is the number of epochs needed before getting the minimal loss on testing set mentioned before.

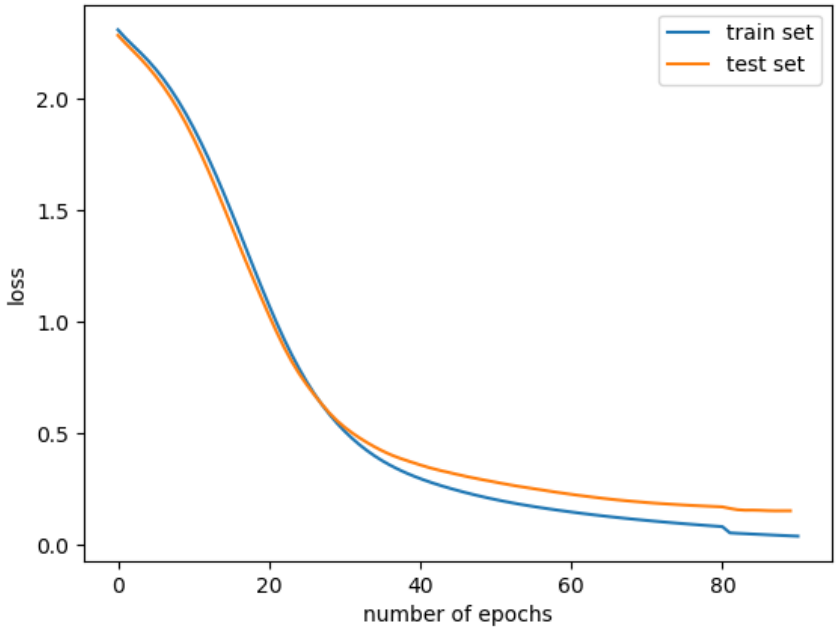

Fig. 10. The trend of loss on training set and test set after applying the BDR method.

In the above figure, outlier removal happens only once during training. It is shown as a jagged falling on the blue line at the 80th epoch. Then, halting training criterion is achieved at the 90th epoch. It means the variance of loss for all instances in the training set is lower than 0.01 at the 90th epoch after outlier removal at 80th epoch.

Accuracy is the key aspect of the results. Thus, we compare the accuracy of the results of our CNN with other methods which perform classification task on the Semeion Handwritten Digist dataset [2].

Table 1. Accuracy of different methods for predicting on the Semeion Handwritten Digits dataset.

	Accuracy
CNN with BDR	93% ~ 96%
CNN without BDR	95% ~ 98%
ML: none	92.66%
ML: PCA	93.50%
ML: All	92.87%
Baseline	92.03%
Auto-WEKA	94.13%

Here is a list of the accuracy of several methods on the Semeion dataset. The accuracy of the methods except for CNN with BDR and CNN without BDR are cited from [3]. Generally, the accuracy of the results of CNN with BDR is better than the methods in [3]. However, what is noticeable is that the accuracy of CNN without BDR is much better. It means the Semion dataset has few outliers and the BDR method [1] cannot improve its quality much. In our prediction, if we reduce the value of variance in the halting training criterion, the accuracy of CNN with BDR will be higher. But it also means the BDR method [1] has fewer impacts on training.

4 Conclusion and Future Work

We finished an experiment of the BDR method [1] on a CNN with the Semeion Handwritten Digits dataset [2]. Although the results show that the method dose not improve the performance much, it helps us explore how the BDR method [1] behave with a dataset containing few outliers.

The BDR method definitely improves efficiency of training and quality of datasets. It also gives the network ample time to learn the feature of those patterns which are candidate of outliers. Combining this with the criterion of halting training, the BDR method has a great advantage in preventing overfitting. However, the effect of the BDR method can be influenced by properties of dataset. In the Semeion Handwritten Digits dataset [2] we picked, the trend of error variance on training set is different from expectation, which promotes the modification on the criterion of starting outlier detection and removal and halting training. The BDR method has few impacts in this experiment. It means the Semeion dataset has few outliers. As mentioned in [1], the BDR method is not suitable for those datasets which is known to be clean as the outlier detection method of the BDR will be not accurate in such datasets.

In the future, we may evaluate the BDR method on more complex real-world datasets or datasets with manual noises. It will help us understand the different performance of the method on different performance and explore its reasons. Especially, we need to explore a solution for the unusual performance on those datasets similar to the ionosphere dataset which have two types of patterns and their proportions are unbalanced. We also need to explore a more generalized rule to determine the criterion of starting outlier detection and removal and halting training. It will improve the stability of the BDR method on different datasets. We should also evaluate those new outlier detection method and compare their results with the BDR method. It will bring a better understanding of the drawbacks of the BDR method and a direction of outcoming them.

References

[1]. Slade, P., & Gedeon, T. D. (1993, June). Bimodal distribution removal. In International Workshop on Artificial Neural Networks (pp. 249-254). Springer, Berlin, Heidelberg.

[2]. Semeion Research Center of Sciences of Communication, via Sersale 117, 00128 Rome, Italy.

[3]. Bürger, F., & Pauli, J. (2015). Automatic Representation and Classifier Optimization for Image-based Object Recognition. In VISAPP (2) (pp. 542-550).