The Basic Research of Data Mining Of Input With Weighted matrix On Data Classification Based on Neural Network

Penghao Jiang

Research School of Computer Science, Australian National University, Australia
u6654495@anu.edu.au

Abstract. How to choose the feature for training the back-propagation neural network is a well-known problem. The basic principles are to avoid encrypting the data's underlying structure and avoid using irrelevant inputs. So people propose to use the data mining technology on the input to solve this problem.There always has some dataset contains many different types of data and real data sets often include many irrelevant or redundant input fields. It will make the model deteriorating performance. So we need to make some operations to change the amount of information. Then we choose the data mining technique to solve this problem. Data mining technique is used for finding problem from the data and solve the data. In this paper, we will search on using the sorting weight matrix of the trained neural network itself whether it can determine which inputs are significant for the network. When we use the neural network to deal with data classification problem, we just need to use the key feature of this data, it means that we need to drop some useless feature of data, If we want to make our model perform better, we need to change the amount of information of the data. This paper use a novel functional analysis of the sorting weight matrix based on a technique developed for determining the behavioural significance of hidden neurons and use BP neural network to work on the data classification problem .So when we use the data mining technique on the data classification problem.The result shows that when we use the weighted matrix technique on dealing the input of the data can get a better performance. The reason of the result maybe is delete some unnecessary information of the input data.

Keywords: Sorting Weight Matrix Technique, Data Classification, BP neural network, Data Mining

1 Introduction

Data classification is a hot topic in today's society. We can see the application of data classification everywhere in our lives. How to improve the accuracy of data classification has become a hot long-term topic in deep learning. Data classification is to merge data with a particular common attribute or characteristic and distinguish it through the feature or aspect of its category. The goal of data classification is to achieve data sharing and improve processing efficiency. It is necessary to follow the agreed classification principles and methods[1]. Data classification has many applications in our society, such as the big data cloud and Spam Classification System.

It means this technique is widely used today, so how to improve the accuracy of the classification becomes a big problem. Using the sorting weight matrix method can delete the unnecessary input

feature, and it also can ensure the accuracy of type. And the sorting weight matrix technique magnitude of the contribution is disentangled from the sign of the contribution. The importance of contributions is significant in indicating whether an input is essential. In this experiment, we can reduce the size of the compression layer and batch size to make the accuracy higher.

This dataset which I use in the experiment includes 120 different feature data. And each kind of feature data has 385 variables. Each feature data represent different kinds of feature data. I should use the feature data to train my network model and make a classification of the data.

This paper mainly uses the sorting weight matrix technique to deal with the input feature data. This experiment will design three different labels to compare the accuracy of the same feature data in different environments. The first environment is the normal model with the n-fold verification technique. The second environment is using the sorting weight matrix technique with the n-fold verification technique. Then make a comparison between their two different environments. The suitability of the sorting weight matrix technique should be determined by comparing their accuracy with these two classification results.

2 Method

2.1 Basic method for classification

I use Backpropagation network to become the basic neural network for classification. The choosing method which I choose to use on the input is sorting weight matrix technique.

2.1.1 Structure of Backpropagation Neural Network

All connections are from units in one level to units in the next level, with no lateral, backward, or multi-layer connections. Each unit is connected to each unit in the preceding layer by a simple weighted link. The network is trained using a training set of input patterns with desired outputs, using the back-propagation of error measures.

2.1.2 Make optimization on the Backpropagation Neural Network

After training the neural network, we can find that the model can not get very high accuracy and low loss. It means that this model can not train well. It also means that some data is underfitting. So I make some optimized operations on the model as follows.

Normalization on Input feature

We can find that the reason for getting worse performance in classification accuracy includes uneven data dimensionality and using interference input data to train the model. Normalization can help us improve the convergence speed of the model. Simultaneously, it can prevent the model parameters from being affected by some very large or minimal data. Then we can find after the normalization of input feature data, and we can get more optimized model parameters with the same input. In this paper, I divide the input data into different subsets, which contain different types of data. The result I showed in Figure 7, Figure 8 and Figure 9.

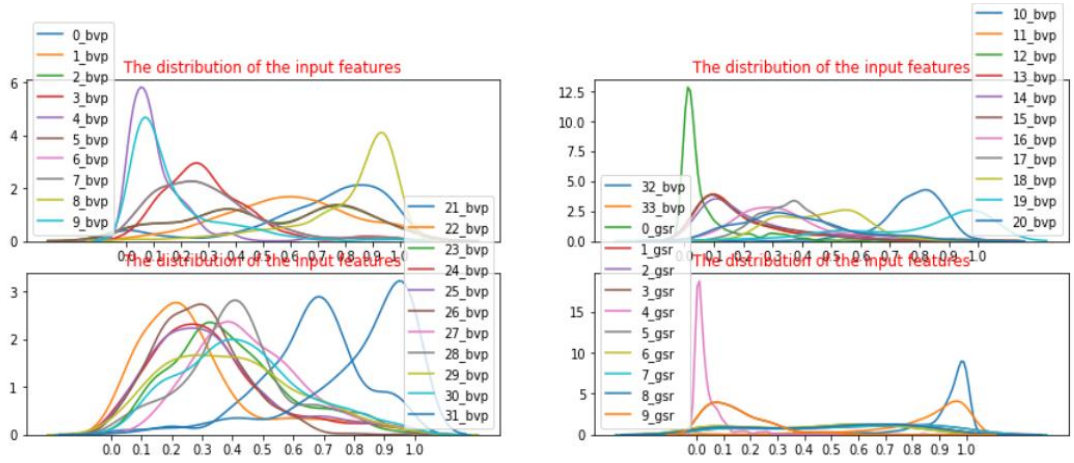

Figure.1. The result of normalization of input data

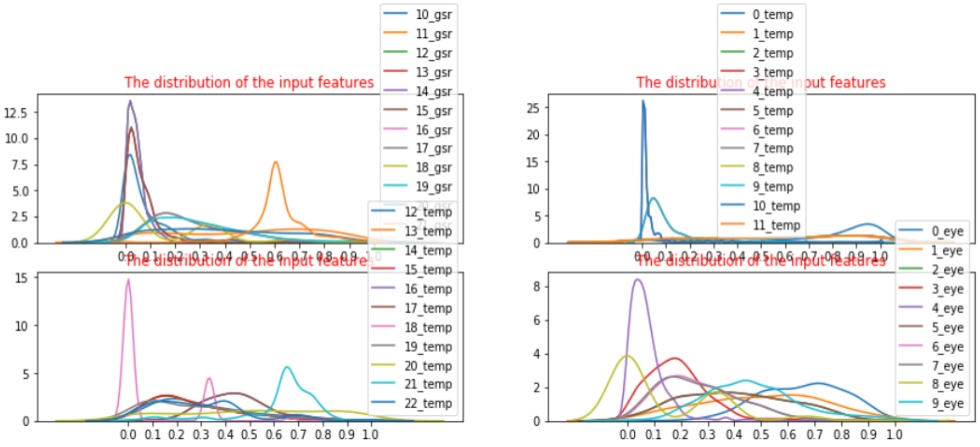

Figure.2. The result of normalization of input data

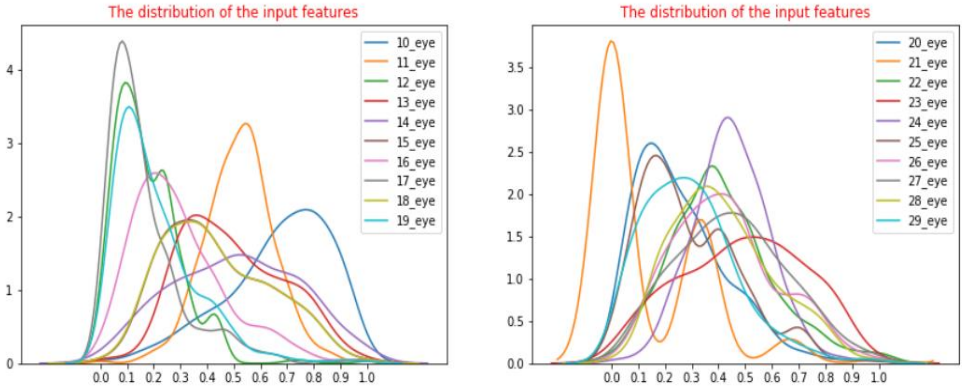

Figure.3. The result of normalization of input data

N-fold Cross Validation structure

I choose to use the N-fold cross-validation technique in my experiment, which means I divide the input data into n subsets. Then I set up a test set for each subset and making other data set to be the

train data. After training the model, run a cross-validation operation for n times. Each time a subset is selected as the test set, and make the average cross-validation recognition accuracy of n times is used as the result of the model. It can help us avoid the limitations and particularities of fixed data sets. And we can test on each data of input data. The technique also divides the data set multiple times and average the results of various evaluations to eliminate the adverse effects caused by unbalanced data division in a single division. Then we can get a model with stronger generalization results and each subset's data distribution average. Another key problem is how to choose the value of n. So I have tried many values of n. I choose n equal 15 for my experiment. If I choose a bigger number for n value, I found that I can train more data for each step, but it will lead to longer training time and lower training efficiency. But it can help the model fit input data more comfortably.

Adam Optimizer Structure

The Adam optimizer algorithm is updating neural network weights repeated based on training data. It designs independent adaptive learning rates for different parameters by calculating the first-order moment estimation and the second-order moment estimation of the gradient. It combines two algorithms in the Adam Optimizer; the first one is the AdaGrad algorithm, the second one is the RMSProp algorithm. The Adam Optimizer inherits advantages from these two algorithms; it can improve calculation efficiency. The Adam Optimizer generate model parameters will not be affected by the gradient size transformation. And there is another problem with which optimizer algorithm to use? I set different optimizer algorithms and show the result in Figure.4.[2]

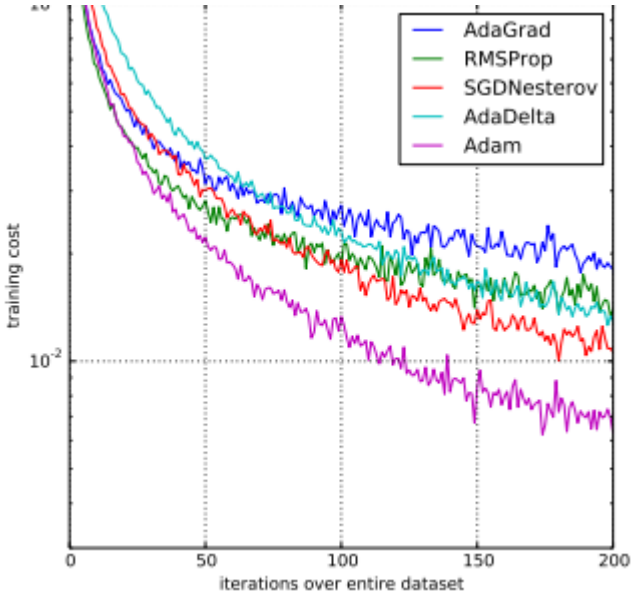

Figure.4. The comparison of different algorithms

Adjustment parameter of model

During the adjustment parameter operation, I find this experiment has many parameter need to change, such as the number of the epoch, learning rate and the number of iterations. Before changing the parameter of our model, I find the model is underfitting. It means the learning ability of the model is

weak or the input feature data is high. The complexity of the model will become more heightened when we add any data with different dimensions. So we need to make the learning rate higher than before; it can make the learning speed of the model faster than before.

2.2 Method for sorting weight matrix of input data

2.21 Structure of sorting weight matrix on input data

In this paper, I use the matrix weight technique on input data proposed by Wong, Gedeon, and Taggart in 1995[3]. I will show this algorithm with formula (1) and formula (2). Then we can get the contribution of an input neuron to an output neuron from these two formulas. To get a higher accuracy for the classification, we need to get more 'useful' feature information from the input[4]. So the normalization (in this paper, I use the sorting weight matrix technique on input data instead). I hope I can get less loss and more accuracy at the same time. So I design three control groups for my experiment. Each group has three different data sets for proving using a sorting weight matrix on input data to delete some inputs that have the lower weight among the information. Then I will show you the process of comparison between these two conditions (one is before using the sorting weight matrix, and the other is after using the sorting weight matrix technique) with Figure 5.

$$P_{jk} = \frac{|W_{jk}|}{\sum_{r=1}^{n} {}^{h} |W_{rk}|}$$

(1)

$$Q_{ik} = \sum_{r=1}^{nh} (P_{ir} \times P_{rk})$$

(2)

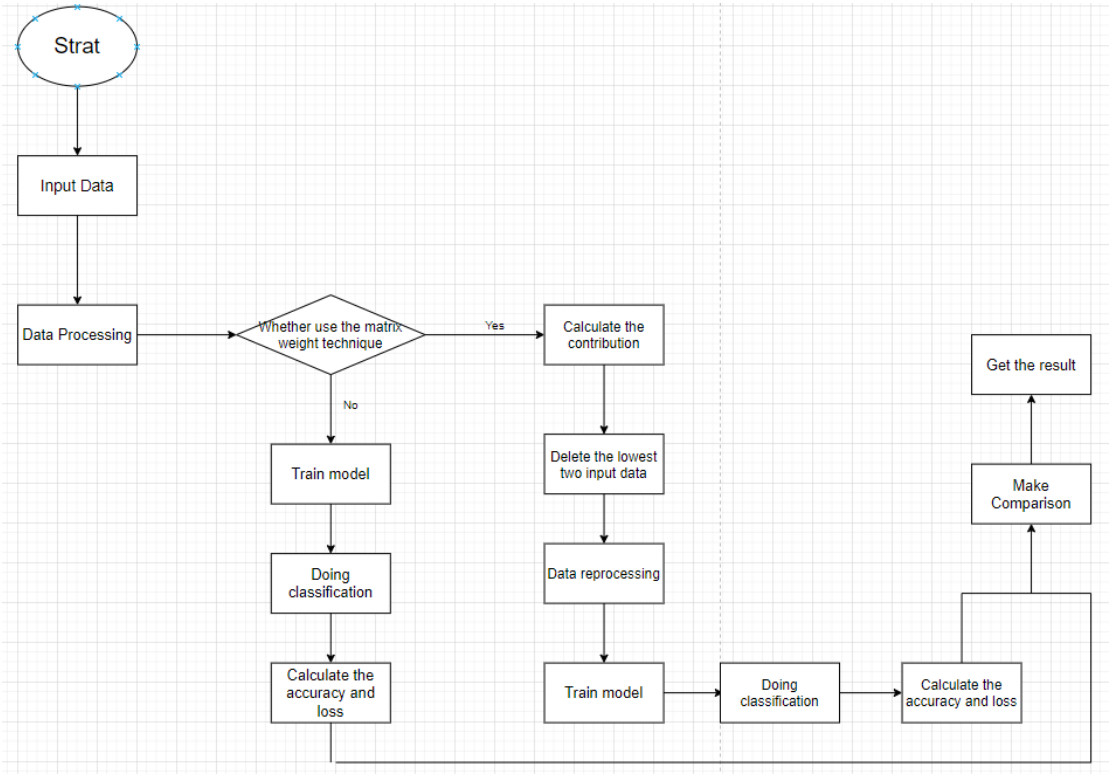

Figure.5. The process of comparison

3 Results of experiment and Discussion on future work

3.1 Result of Optimization

After using the optimizer, which is shown in 2.1.2 on the Backpropagation neural network, the comparison of the result of three labels is shown with Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11 below. Figure 6, Figure 7, and Figure 8 show us the result of the loss of the model, and the other Figures are displayed for the result of model accuracy. When we use the sorting weight matrix on the input data, the model's training loss and test loss are reduced significantly. The accuracy of classification also has improved than before. It proves that the sorting weight matrix technique is effective in making the model perform better. It can delete unnecessary input data to make the model fit better than before. The selection input deletion is performed by calculating the contribution ranking by the sorting weight matrix technique is useful. This measure can effectively delete the non-representative feature input, thereby reducing the loss of the model and increasing the accuracy.

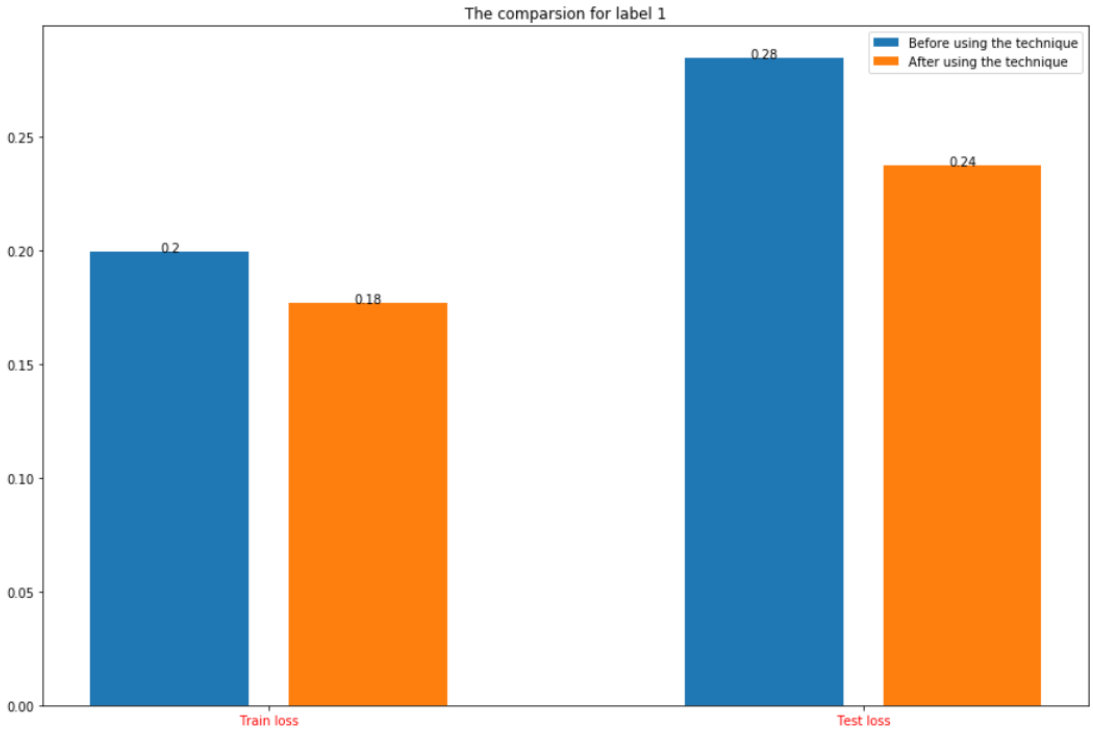

Figure.6. Train loss and test loss for the label 1

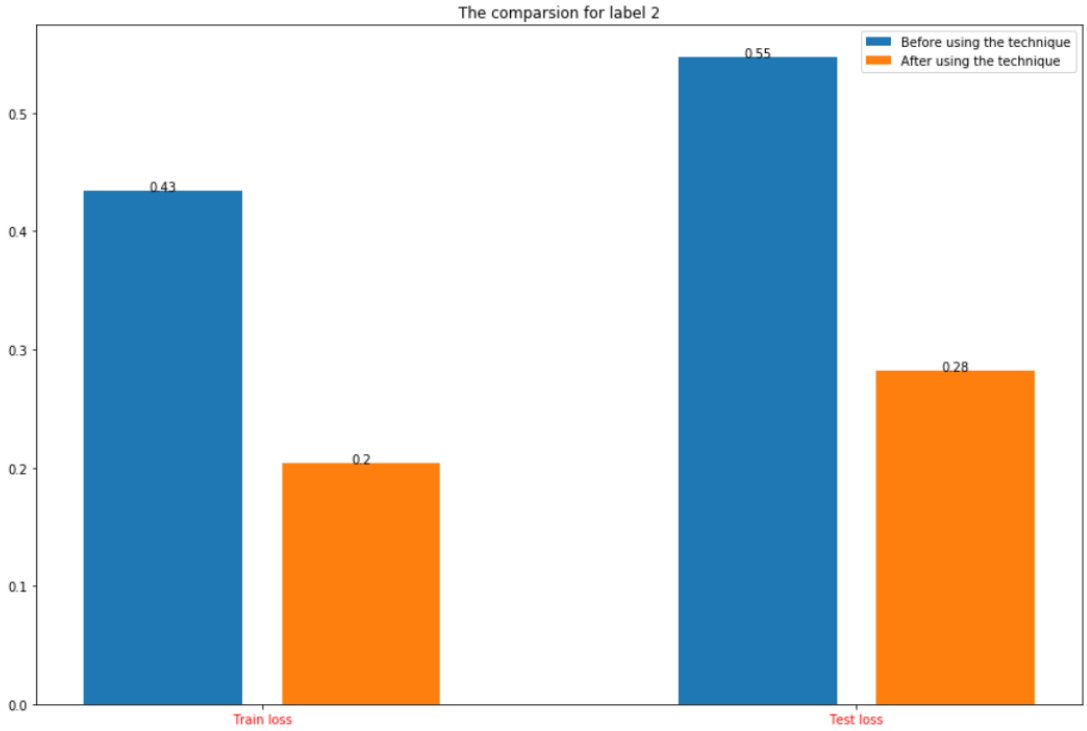

Figure.7. Train loss and test loss for the label 2

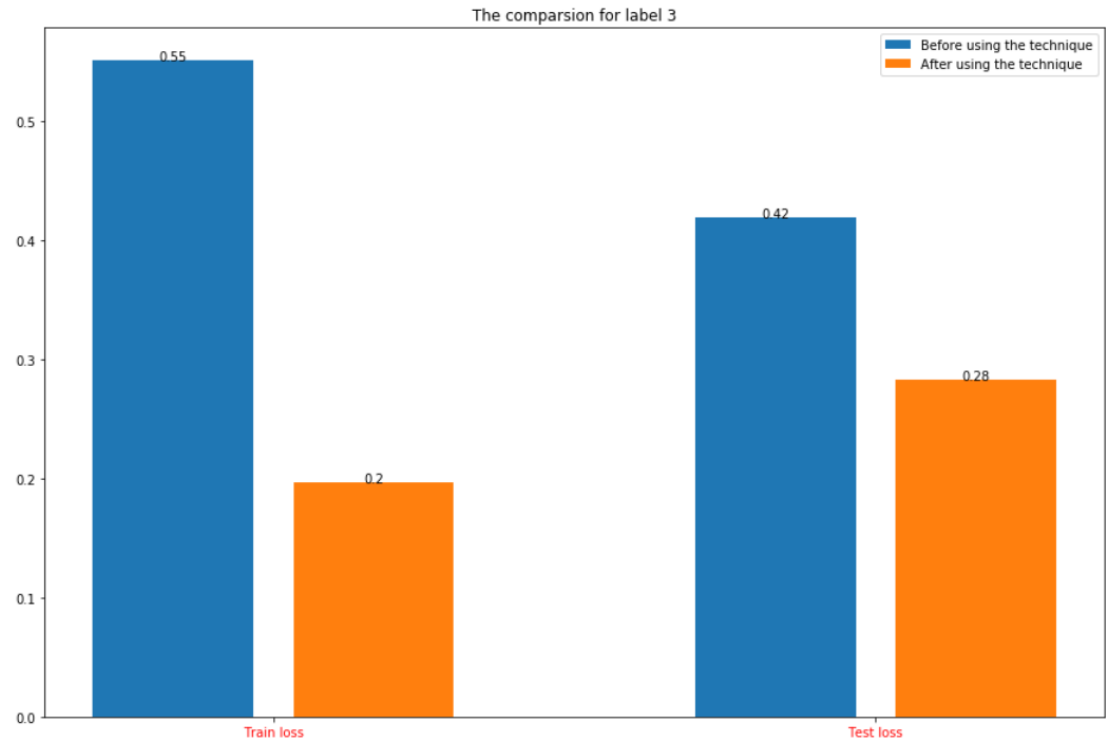

Figure .8. Train loss and test loss for the label 3

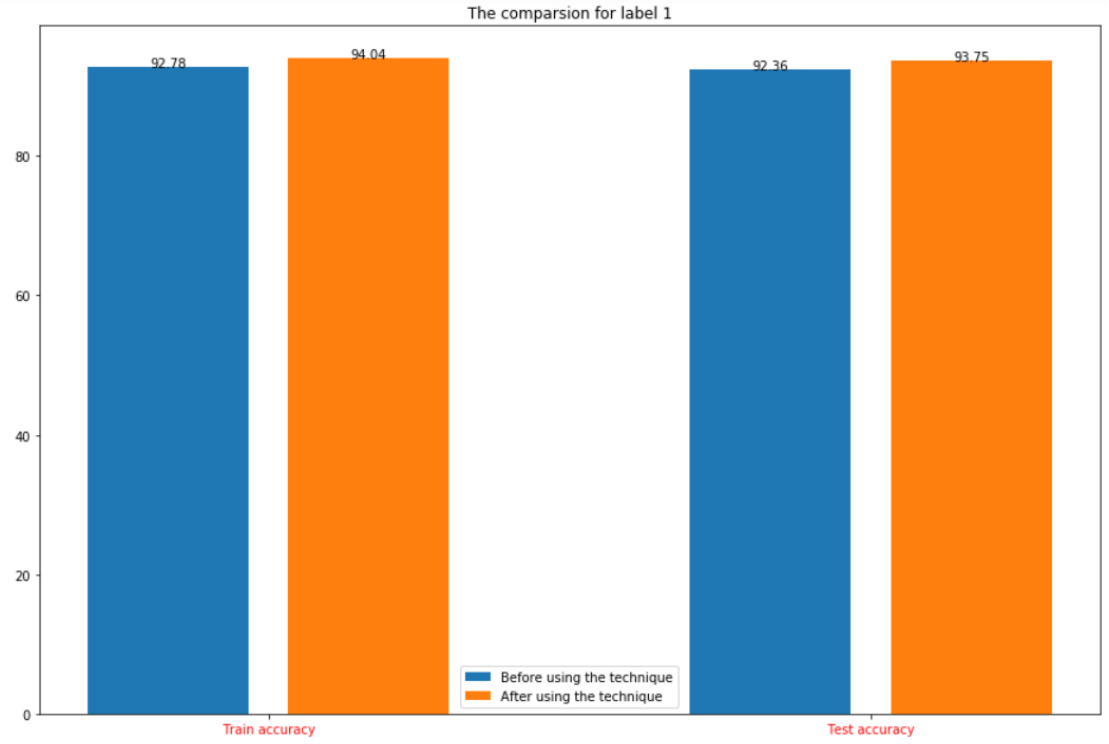

Figure .9. Train accuracy and test accuracy for classification label 1

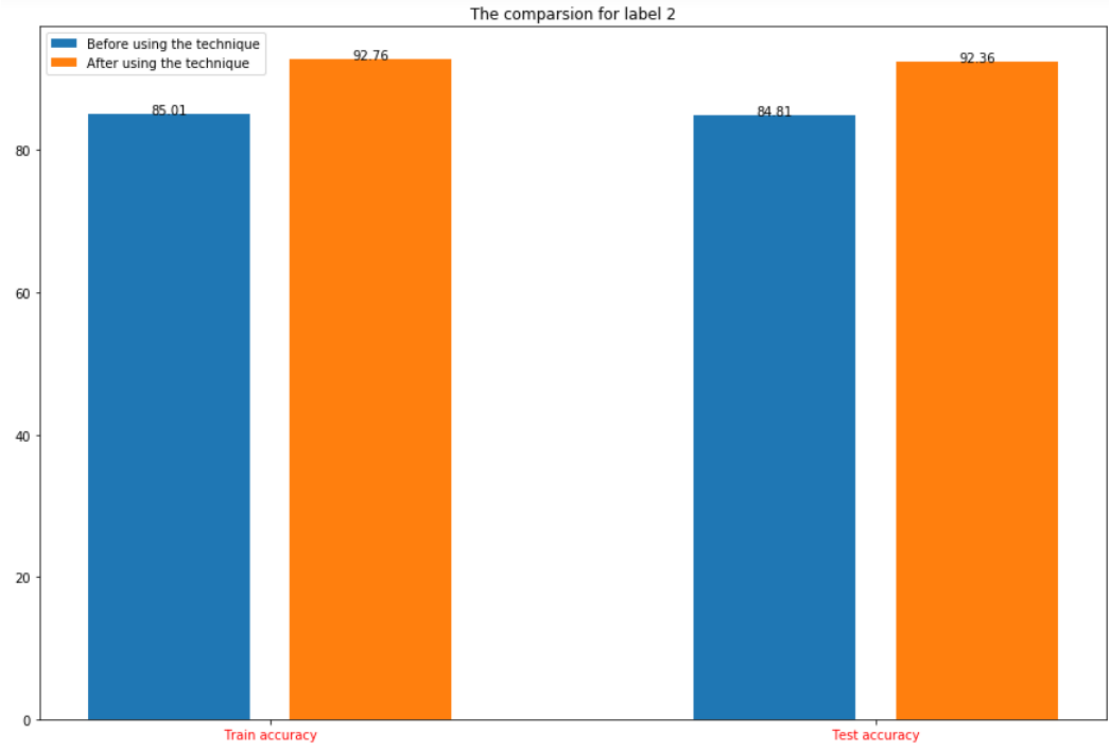

Figure .10. Train accuracy and test accuracy for classification label 2

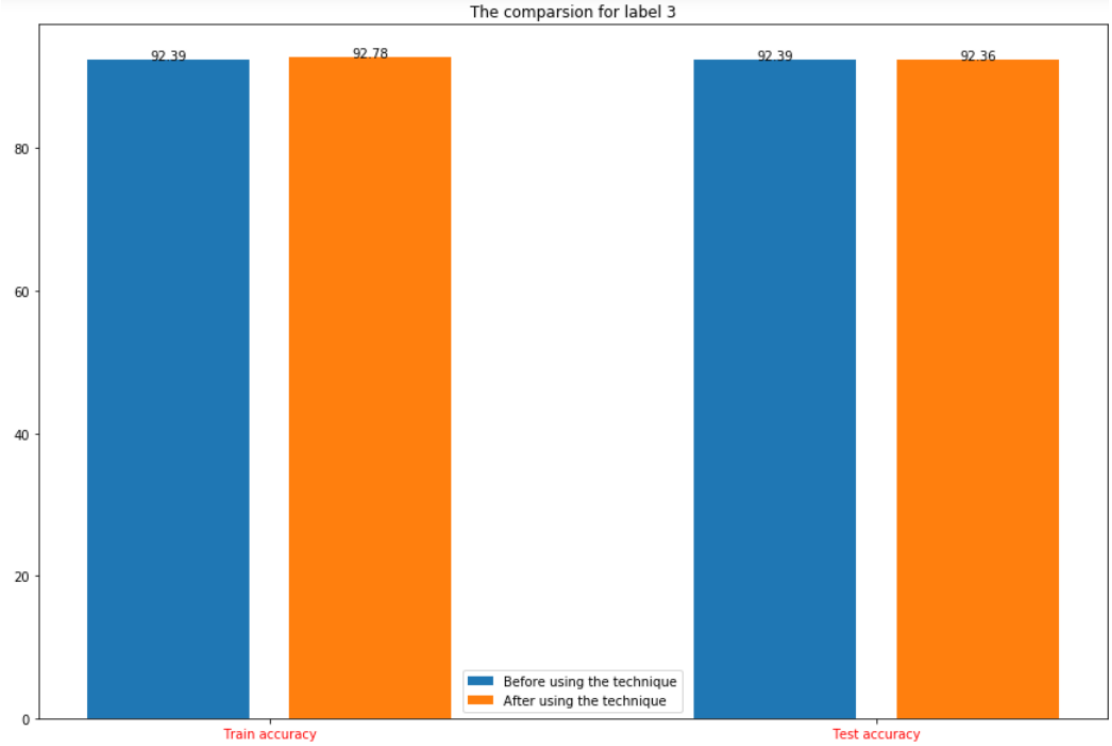

Figure .11. Train accuracy and test accuracy for classification label 3

3.2 Result of the experiment and future work

From the Figures above, we can find the classification test accuracy can get almost 93%. I have repeatedly do this experiment, and I find the accuracy and lost of the model are stable. The lost of the

model from 0.55 reduces to 0.2, and the accuracy increase from 92 to 93.5. We can find that the accuracy rate has been improved to a proper degree, and the lost has been dramatically reduced. And I think there are still have some data that can not be fully fitted in the model, which is why the accuracy rate cannot be greatly improved. I will try to solve this problem in the future work.

I will try to use another formula to calculate the weight of input and combine the technique with another model type in future work. And I will use the higher dimensional image for the input data then repeat the experiment to get the loss and accuracy to evaluate the future work. The second future work point is using other dataset to observe the result whether better or worse, if we can get the better result in other dataset, it proves the sorting weight matrix technique is useful to make our model perform better.

References

1. Gedeon TD. Data mining of inputs: analysing magnitude and functional measures. Int J Neural Syst. 1997 Apr;8(2):209-18. doi: 10.1142/s0129065797000227. PMID: 9327276.
2. SiYuan. https://www.jiqizhixin.com/articles/2017-07-12, The heart of machine learning, 2017.
3. Wong, PM, Gedeon, TD and Taggart, IJ "An Improved Technique in Porosity Prediction: A Neural Network Approach," IEEE Transactions on Geoscience and Remote Sensing, vol. 33, n. 4, pp. 971-980, 1995.
4. Dmitrii Borkin,Andrea Peterkova Nemethova,German Michalconok, September 2019 Research Papers Faculty of Materials Science and Technology Slovak University of Technology 27(45):79-84