Basic Research On Data Classification With Sensitivity Analysis

Penghao Jiang[1]

Australian National University, Canberra, Austrlia u6654495@anu.edu.au

Abstract. In recent years, with the rapid development of neural network technology. People's cognition of neural networks is becoming deeper and deeper. Through research, scholars have discovered that people can classify and predict data through different models. Furthermore, the essence of the neural network is which can derive conclusions from a complex and seemingly unrelated set of information [1]. There will be many influencing factors in the forecasting process. So we have to analyze clearly which factors are dominant and which factors have a relatively small influence. It requires us to conduct sensitivity analysis on the characteristic elements to solve the problems we encountered during the analysis.This paper will use a feed-forward neural network with three layers to train the data from the subjective'-' belief '-'observers"-' features'-' labels.csv. The purpose of using neural networks is to predict the subjective information provided by the speaker from the various data which we get from the dataset detected from the audience. In order to solve the proposed hypothesis, I will normalize the data to observe the contribution of each data in the data set. Secondly, the paper will use the PCA algorithm and the decision tree model[2] to verify the hypothesis and calculate the maximum Sensitivity analysis with max gradient as the boundary to assist verification. The last step is making a comparison between our network with other network models.

Keywords: neural network · PCA algorithm · Sensitivity analysis method · rule extraction theory · decision tree model

1 Introduction

1.1 Motivation

Data classification has become a hot research topic in neural networks, how to improve the performance of neural networks in the process has become a vital issue in the field. In this paper, we can find that a well-trained neural network can predict the data, and we can also get the accuracy and prediction value from the network. However, we cannot know how the neural network classifies and weights the input data. From this paper we can find by preprocessing the data and using the sensitivity analysis algorithm on the input data set, it is possible to obtain which value of each feature attribute has the more significant effect in operation and which value has the more negligible effect.

When training neural networks, we often encounter situations where the data set is not large enough to cause overfitting. The essence of overfitting is to have unknowingly extracted some of the residual variation (i.e. the noise) as if that variation represented the underlying model structure. However, it is difficult for us to judge whether the model has overfitting. If we use rule extraction theory in neural networks, we will find that we will get an inferior result when there is overfitting[3]. We also found that when the neural network has over-fitting, we can analyze the input data through the sensitivity analysis method to process the data to reduce over-fitting. Therefore, the rule extraction theory and the sensitivity analysis algorithm are critical to improving neural networks' realization.

1.2 Introduction of Dataset

We can find that the detection system will detect abnormal data when the observed person lies in the experiment through research[4]. The abnormal data of the liar will have a subtle influence on his behaviour, tone, and movement posture. Therefore, we use neural networks to analyze the data and physiological responses of the observed person and predict the results. To proves whether some particular information is controlled. This data set uses the BVP and GSR equivalent values obtained by 23 participants in the experiment on 16 videos randomly selected from 32 videos.

1.3 Sensitivity Analysis Method

After reconstructing the data set, we need to explore the impact of each input feature variable on the output[5]. So we have to perform sensitivity analysis on the input data to get the influence of each input feature data on the output. The output gradient value determines the sensitivity value to the input feature data.

2 Method

2.1 Basic Method of Neural Network

The paper will use feed-forward network to become the basic network. Then I choose the rules extraction algorithm and sensitivity analysis method to become the auxiliary data processing method.

2.2 Normalization on Input Data

We can find that there are not missing-value in the dataset. So we do not need to make operation on filling data. And I find the distribution of different parts of data are different. 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 performance of the model with using the same data set.I show the feature data distribution with Figure1,Figure2 and Figure3.From the distribution Figure, It is not difficult to find a significant difference between the maximum variable and the minimum variable. It may cause the neural network to learn larger values and ignore smaller values. However, for a well-performing neural network it should not be affected by the magnitude of the value. So we should make a normalization on the data set.[6]

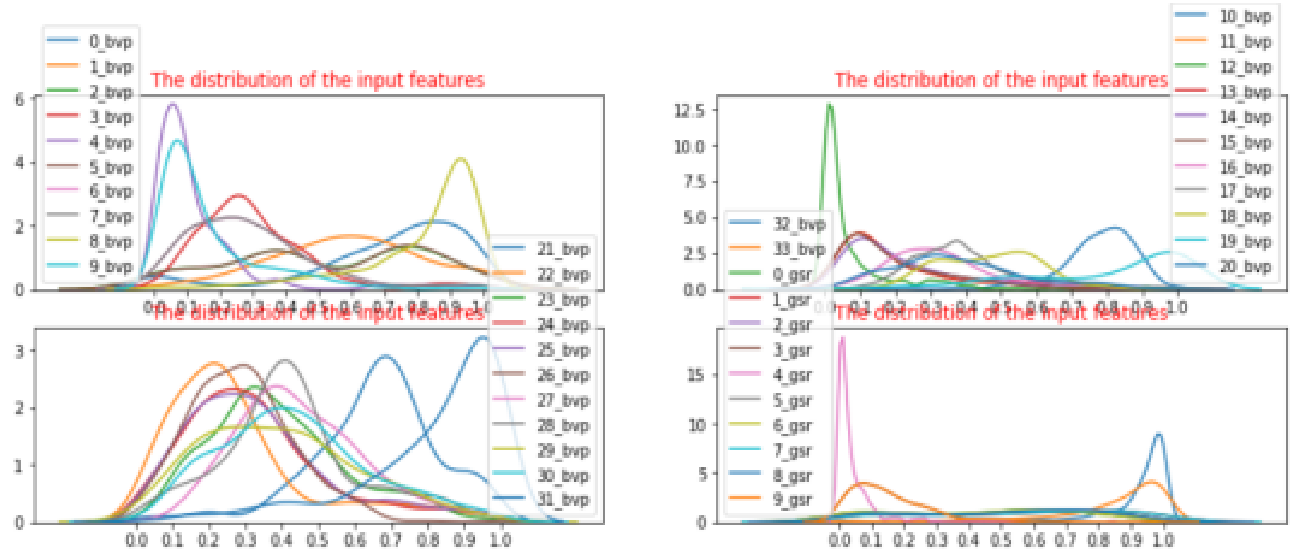

Fig. 1. The distribution of each data in the dataset.

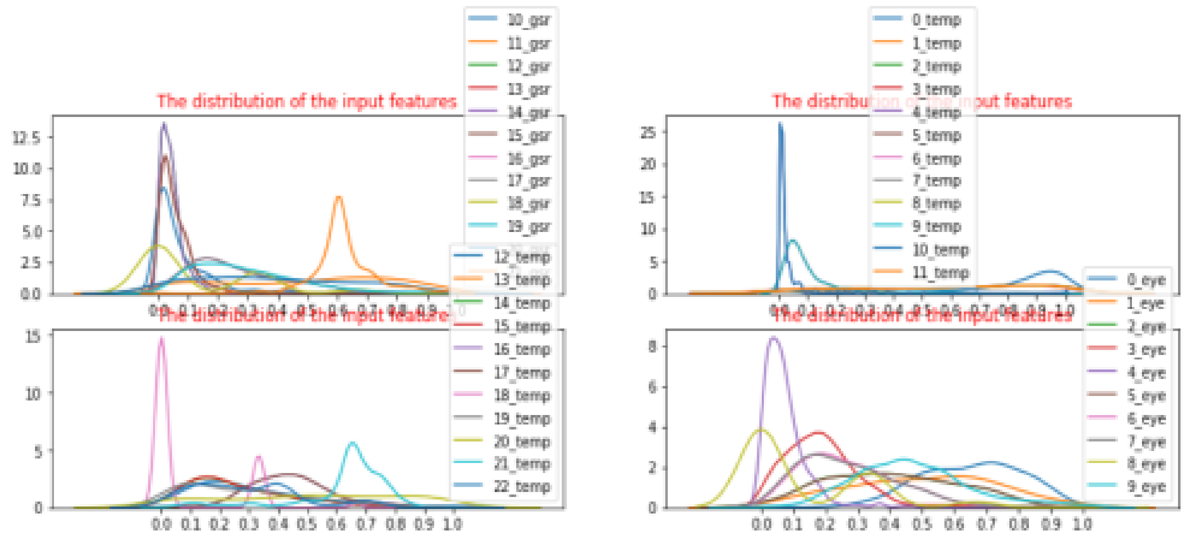

Fig. 2. The distribution of each data in the dataset.

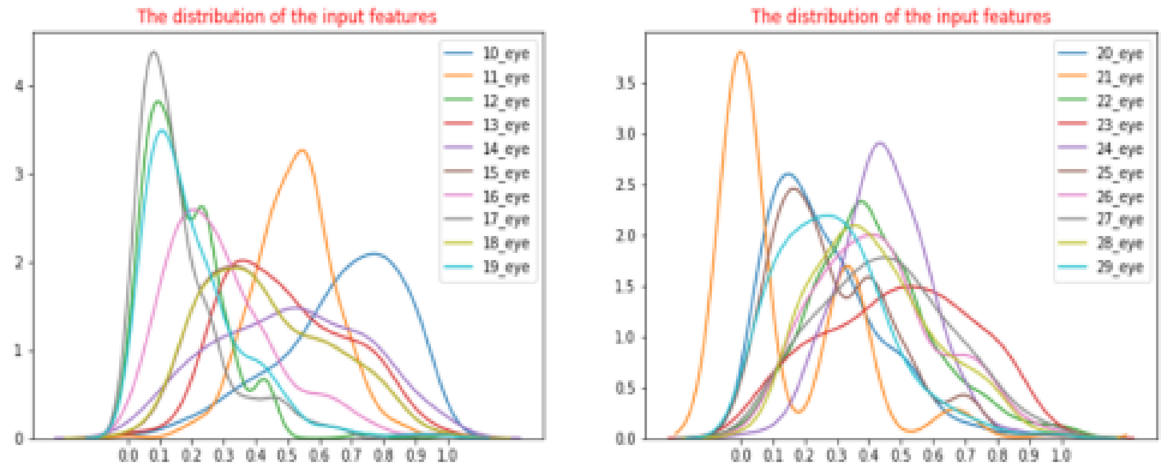

Fig. 3. The distribution of each data in the dataset.

2.3 Operation on Dataset

First of all, I make a drop operation on the first column of data in the data set because the first column of data is only the id value. It will not contribute to predicting the neural network, and the predicted result does not include the id value. In the second step, the abnormal data will be removed by calculating the standard deviation of each variable. Furthermore, the network set the standard deviation value greater than four as abnormal data. We can find that this data set contains four different data types, namely BVP, GSR, TEMP and EYE. Each data category includes 32 minor feature attributes. It is not appropriate to directly use 32 feature attribute values to input the neural network. So I will calculate the maximum, minimum, average, standard deviation, variance, the value of means of the absolute values of the first difference, and the value means of the absolute values of the first difference of four different data types to reconstruct the data set.

2.4 Division of Data Set

I will divide the data set into three parts. The first part is traing set which use 60% data. The second part is valid set which uses 20% dara. The last part is test set which use the last 20% data. Then I use the random state parameter to control random model.

2.5 Select Feature Method

We need to classify the input feature values through the output results. Whenever we input new data to the neural network, it will automatically calculate the distance to the centre of other clusters. Then select the closest cluster to determine the input model.

2.6 K-fold Cross Validation Structure

I choose to use the k-fold cross-validation technique[7] 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, the network will 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.

2.7 Adam Optimizer Structure

The Adam optimizer algorithm is updating neural network weights repeated based on training data[8].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

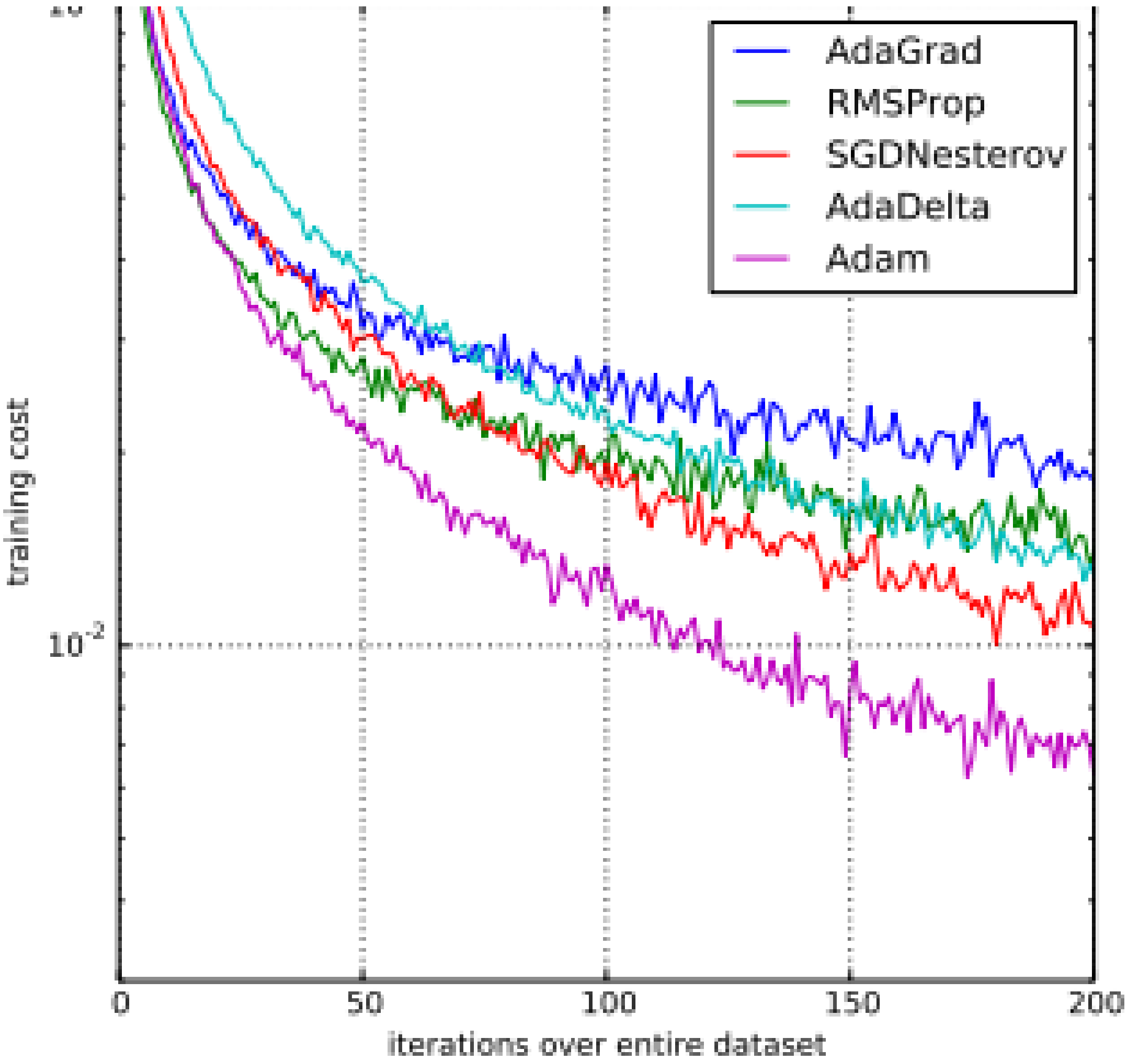

Fig. 4. The comparison of different optimizer method.

2.8 Details of Sensitivity Analysis Method

We can find if the function can get a big derivation value of one data point, which means that the speed of change is faster at the data point. At the same time, we can assume that a data set of any continuous variables named var is entered, then we can call this data set Dataset. We can get from part1.3 that we need to calculate the gradient value from output value to the input. We can regard the gradient value as a decision boundary. We can name this decision boundary as bound. We can assume that any continuous variable var is the decision boundary of the attribute. Any small change in this decision boundary will cause a big change in the output. We can reduce the sensitivity analysis method into finding the largest first-order derivative value for the input feature variable through the derivation of the above two parts. Through experiments, we can speculate that the boundary of the data is allocated by the contribution of the sensitivity of the input feature data for any feature data. So in this method, we should find the maximum value of the first derivative first.

After obtaining the maximum value of the first derivative of the characteristic data, the next issue that needs to be paid attention to is how to trigger the boundary. To solve this problem, I set two parameters, we can name them c1 and c2, c1 is used to count the number of times less than the boundary value, and c2 is used to count the number of times greater than the boundary value. If c1 is greater than c2, we output the corresponding output type. If c2 is greater than c1, the cluster with the most selection of variables greater than the boundary value is the output.

2.9 Evaluation Function

I will divide the evaluation equation into two steps. The first step is to classify the variables. The second step is to compare with the decision boundary. In the classification process, we use the k-means model to classify the input feature data. Then we can get each cluster of the feature data. Then compare with the decision boundary, if it is greater than the decision boundary, the output will change to a new cluster. Otherwise, directly output the obtained cluster.

2.10 Finding Decision Boundary Function

When I am looking for the decision boundary, I tried a variety of different models to test and get each accuracy. When we are calculating maximum value, we can use polynomials instead of approximation formulas. However, how to determine the highest power of a polynomial has become a new problem. So I have tested different models. Moreover, I choose second-degree polynomials and third-degree polynomials. However, the effect is relatively general. So I used the curve fitting function to fit the function. The fitting process is shown in formula 1, 2, and 3. Thus, the offset value between the polynomial function and the fitting function is obtained. The goal of our calculation is to minimize the offset value.It means to minimize the output value of formula 3 . Find the maximum value in the gradient matrix of each model, and I show the results of the test through table1.

$$y(x, w) = w_0 + w_1 * x + ... + w_m * x^m = \sum_{j=1}^{M} w_j * x^j \tag{1}$$

$$Result^2 = \sum_{n=1}^{N} [y(x_n, w) - f(x, y)] \tag{2}$$

$$\delta(Result^2)/\delta(w_i) \tag{3}$$

2.11 Local Regression Function

From table1, we can find that we have used the local regression model for testing. Because in some cases, linear regression does not fit the data well to make good predictions. Because it easily leads to under-fitting. So we use local regression for comparison. The local regression model allows to introduce some deviations in the estimation and assign weight to each point simultaneously. Each data point has its specific window limit. In each window, the closer the data point to the centre, the greater the weight, and vice versa.Then I need to design the weight matrix.

Table 1. The result of different accuracy.

Model type	test accuracy
Decision Tree model	**95.38%**
NearestCentroid model	**63.07%**
ExtraTree model	**93.38%**
RandomTree model	**92.08%**
local regression model	**73.84%.**
Maximum value model	**53.15%**
Square polynomial model	**51.26%**
cubic polynomial model	**52.71%**

The essence of the local regression algorithm is the least weighted square method. For any fitting point, the value of nearby points has a more significant influence on the fitting line. The model need to consider the weight matrix model. It will use tricube weight function as the weight equation. The process is shown in formula 4. After deisgning the weight formula, we need to use the weight formula to calculate each data in the window range.

$$W(u) = (1 - u^3)^3 \tag{4}$$

2.12 PCA Method

It can be found from parts 2.2 and 2.3 that 32 decision boundaries can be obtained by the reconstructed input data set. It means that each row of feature data has a corresponding decision boundary. At the same time, we can find that if there is a particular column of data that does not have much influence on the result through analysis, but this column of data still exists in the input data, it will affect the accurate value of the data prediction. Therefore, we need to use the PCA method. PCA transforms the original data into a set of linearly independent representations of each dimension through linear transformation, which can be used to extract the main feature components of the data. This method often used for dimensionality reduction of high-dimensional data[9]. And I test the accuracy on different model with PCA method. The result I show with Table2, 3 and 4.

Table 2. PCA on different model.

model type	Remaining dimensions number equals 0	Remaining dimensions number equals 5
local regression model	**40.15%**	**59.48%**
square Polynomial model	**40.75%**	**51.28%**
cubic Polynomial model	**39.45%**	**49.23%**
maximunm value model	**39.12%**	**54.33%**

Table 3. PCA on different model.

model type	Remaining dimensions number equals 10	Remaining dimensions number equals 15
local regression model	**56.58%**	**69.94%**
square Polynomial model	**61.27%**	**53.35%**
cubic Polynomial model	**57.65%**	**53.22%**
maximunm value model	**52.18%**	**63.74%**

3 Results and Discussion

3.1 Result of Different Models

We can find that the local regression model has the highest accuracy rate of 63.07% among all the tested models from the result table above. Compared with other models, the local regression model can balance the deviation and

Table 4. PCA on different model.

model type	Remaining dimensions number equals 20	Remaining dimensions number equals 25
local regression model	**65.36%**	**63.74%**
square Polynomial model	**58.39%**	**58.67%**
cubic Polynomial model	**59.63%**	**58.88%**
maximunm value model	**68.72%**	**68.16%**

variance values cleverly. At the same time, data within a range can be extracted as input, and the data within the fitted range will continue to advance[10]. Other models perform fitting operations based on all input data, but the most consistent result is often not always a curve but requires multiple fitting curves to be spliced. Moreover, local regression is not sensitive to the outlier. That is why the local regression model can obtain the highest accuracy.
We can find that the input data set has 32 rows, and other models are difficult to model for nonlinear data or polynomial regression with correlations between data features. It isn't easy to express highly complex data well. Therefore, a lower accuracy result will be obtained.

3.2 Result of Making Comparison With Different Predictive Models

We can find the best performance among all models is the decision tree model through the data table1, with an accuracy rate of 95.38%. The Extra Tree model is closely followed, with an accuracy rate of 93.38%. Moreover, the accuracy rate of the decision tree model is far greater than the local regression model. It can prove that the performance of the decision tree model is better than that of the sensitivity analysis algorithm. The first reason is that the decision tree structure is simpler than other models. The decision tree model does not require too much data input to get a good performance. In addition, the decision tree model does not have many conditions of input data.[11] It also can handle irrelevant feature data well. At the same time, we can find the irrelevant feature data will harm predictions. So this is one of the reasons for the excellent performance of the decision tree model.
The second reason is that the number of input feature data is not enough for perfect polynomial fitting. When the amount of input feature data is not enough to predict perfect, the fitting performance will not be good enough. Furthermore, the sensitivity analysis algorithm calculates the gradient value every time. It consumes much longer than the decision tree model. Therefore, the performance of the decision tree model is much better than the sensitivity analysis algorithm.
However, the decision tree model also has shortcomings. The decision tree model has the problem of poor generalization performance. At the same time, small data changes may lead to big changes in the model, leading to huge deviations in the prediction results. However, the sensitivity analysis method uses a local regression method, which can effectively improve generalization. For data including categorical variables with different numbers of levels, information gain in decision trees is biased in favour of attributes with more levels.[2]

3.3 The Optimizer Method Choose

In this paper, I choose the Adam optimizer. However, we have tried many other optimizers in this paper. All of them get worse performance than Adam optimizer. In this part, I will analyze the reason why we choose the Adam optimizer. From Figure 4 we can find that as the iterations number of the entire dataset runs increases, the training cost of Adam Optimizer drops the fastest. Second, if we choose the SGD optimizer, how to choose a perfect learning rate is the problem of the optimizer. If we choose a small learning rate, the convergence speed will very slow. However, if we choose a big learning rate, the loss function will keep oscillating around the minimum value. The SGD optimizer may also oscillate at the saddle point at the same time. As we know, the Adam optimizer can divide into two parts. The first is the momentum method, and the second is the RMSProp method. The learning rate of the Adam optimizer can adaptive reduction and ensure that the iteration is relatively stable.

4 Conclusion and Future work

4.1 Conclusion of Paper Result

At the beginning of the paper, a hypothesis about the behaviour of hidden neurons in neural networks was proposed. In this paper, I first performed normalization and reconstruction processing on the input data set. Afterwards, the data is processed for dimensionality reduction through the PCA method. When the model using the PCA method, I have tried to change different fitting models to test simultaneously. Furthermore, use multiple models (such as decision trees, extra tree model, etc.) for individual testing. Then select the best model to compare the best result(local regression model).We found that the data set processed by the weight matrix alone cannot effectively improve the performance of the model, so we adopted the sensitivity analysis method to replace the weight ranking method. Although the accuracy of the decision tree model is higher than local regression, the sensitivity analysis method using local regression is more stable than decision tree model. Furthermore, it shows the composition and results of the neural network very well.

4.2 Future Work

In future work, I will try to improve the stability of the decision tree model. At the same time, I will try to solve the problem that the depth of the decision tree model is too large, which causes too many splits times. Then I propose to get a new calculation formula to calculate the gradient value in the sensitivity analysis method to reduce the run time. And it will also need a new fitting optimization method for the poor fitting effect caused by insufficient data. For example, the partial derivative of each part of the fitting is equal to 0 to solve. Alternatively, use the gradient descent equation to optimize the fitting equation.

From the point of view of the neural network component itself, this paper uses a three-layer network. Too few layers lead to insufficient network expression and weak learning ability. It will destroy the performance of the model. We will try to use the LSTM model or neural network with more layers for training in future work. Then combine with the sensitivity analysis method and compare with the results of this paper.

We can find that the speed of the Adam optimizer will sometimes slow because of calculating bias correction. In future work, we will try to set an accelerated part in the optimizer. Such as we can consider the ”future location” in calculating the gradient of the next step. Then we can use the momentum of the next step to calculate in the previous iteration. It means correcting when the gradient is updated to avoid moving too fast and increase the sensitivity at the same time. We can use the new momentum instead of the traditional momentum without bias correction. It will reduce the training time.

References

1. ”Neural Net or Neural Network - Gartner IT Glossary”, www.gartner.com.
2. Deng,H,; Runger, G;Tuv, E. (2011): Bias of importance measures for multi-valued attributes and solutions In: Artificial Neural Networks and Machine Learning, ICANN 2011 - 21st International Conference on Artificial Neural Networks,Proceedings (PART 2 ed., pp. 293-300), LNCS, vol. 6792, No. part2.https://doi.org//10.1007/978-3-642-21738-838
3. Towell, G.G., Shavlik, J.W. ”Extracting refined rules from knowledge-based neural networks”. Mach Learn 13, 71–101 (1993). https://doi.org//10.1007/BF00993103
4. T. D. Gedeon and H. S. Turner, ”Explaining student grades predicted by a neural network,” Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan),1993, pp. 609-612 vol.1, https://doi.org//10.1109/IJCNN.1993.713989
5. Gedeon TD. Data mining of inputs: analysing magnitude and functional measures. Int J Neural Syst. 1997 Apr;8(2):209-18. https://doi.org//10.1142/s0129065797000227. PMID: 9327276.
6. ”How, When, and Why Should You Normalize / Standardize / Rescale Your Data?”, https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff.
7. ”A Gentle Introduction to k-fold Cross-Validation”, https://machinelearningmastery.com/k-fold-cross-validation/.
8. Kingma, Diederik; Ba, Jimmy (2014).”Adam: A Method for Stochastic Optimization”. arXiv:1412.6980
9. Bengio, Y.; et al. (2013).”Representation Learning: A Review and New Perspectives”. IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (8): 1798–1828. arXiv:1206.5538. https://doi.org/10.1109/TPAMI.2013.50
10. Harrell, Frank E. , Jr. (2015) Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer. ISBN 978-3-319-19425-7.
11. Gareth, James; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2015).” An Introduction to Statistical Learnings”. New York: Springer. pp. 315. ISBN 978-1-4614-7137-0.
12. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
13. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)