Analysis of ionosphere data and improvement

Huiwen Zhang

Canberra ACT 2601
u6342618@anu.edu.au

Abstract. In this report, first, a junior neural network was established to finish a simple prediction about the data, which is radar data to tell whether the radar is good or bad. And then some little improvements about the data size and the input was made in the program. However, though there are some improvements, the result of prediction is still worse than the result of prediction in the published research paper. So there are still some aspects, like algorithm and training methods that it can make a progress in this prediction. And in this report, a genetic algorithm and LSTM are used to improve the prediction. The genetic algorithm is used to make the LSTM better by improving the parameters and the LSTM help to deal with the data with a more complex process.

1 Introduction

The first problem is a binary classification task. There are 34 features in this dataset and all these features are continues. All these features are used to classify the radar whether a good one or not. Here a basic neural network and a more specified method on input are used to improve the accuracy of the result. However, the result of the improvement is not so obvious and there are still many methods can be chosen in this case to reach the accuracy in the research report. And in this report, genetic algorithm helps to solve a the most suitable parameters to give into the LSTM, and then the LSTM will help to deal with the data to give a prediction. At this time, the loss will be used as the parameter to display the degree of the suitable of the parameters when doing LSTM training and testing.

1.1 data chosen

The reason why this data set is chosen is that those data are all consisted of numbers, so it does not need to be preprocessed to change those data into numbers. And there are also no missing numbers in those data so that it is not necessary to remove those missing value or fill some value in. Third, there are 34 features in this data set and it is suitable to study the amount and meaning of features if can affect the final result of prediction. Forth, there are only more than 500 lines so that it can save a lot of time to train and test. In conclusion, this data set is convenient enough to use directly and not so big to predict fast.

1.2 the problem

The problem of this data set is to use the features which are from 17 different pulse numbers for the Goose Bay system to classify if the radar is good. All the 34 features are continuous, and the method is classification. The final class is to tell the radar good or bad. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. This is a binary classification task. This time, the problem is that how we can use the genetic algorithm to help to deal with the choose of the parameters which will be used in the LSTM. And comparing the loss between different conclusions to see if the genetic algorithm indeed help to select the best combination of parameters.

1.3 the outline

First, it should be investigated clearly that the meaning of each feature of this data set and the final aim that the classification will do. Second, a basic neural network was established to make a classification about if the radar is good. The data are divided into two parts, the training set and the test set. The neural network was used to train the training data set first and then to do the classification in the test set. And then we can compare the prediction and the original data to get the accuracy that we predict. After reading the paper, some features are changed to fewer and some input are changed using the encoding methods. Then the accuracy of this model can be improved a little, though the result is not as good as the research paper. So there are still some method can be tried to improve the result. In this report, a genetic algorithm was established to select the parameter, and then the best parameter were given to the LSTM algorithm to train and test the data set. The parameter used to see which parameter is better is loss. If the loss is smaller, then the conclusion is better.

2 Method

2.1 the data

First, the three layers neural network are used to do the classification. There are 34 inputs and 1 output in this model. Because the data set is not so complex and there are no missing values, so there is a very simple preprocess to clean the data. And the result of the data set is 'g' and 'b' which represents good and bad. So it should be turned into 1 and 0 to be convenient to do the classification.

2.2 the basic neural network

The basic neural network is a three layers neural network with 34 inputs, 1 output and 100 hidden neural. It includes five steps: download and import all required libraries, load and setup training dataset, define and train a neural network, load and setup testing dataset and test the neural network. There are some details in this neural network. The amount of hidden neural and the number of epoch are all have deep influence on the accuracy of the final result. Too much or too little of them may cause a huge difference in accuracy. The table below (Table 1) show that the accuracy in different condition when finding the most suitable condition for this model.

Table 1. accuracy of original model

Hidden neural	Epoch	Accuracy
80	500	64.76%
	700	60.95%
	900	65.71%
	1100	79.05%
	1300	80.95%
	1500	83.81%
100	500	63.81%
	700	83.81%
	900	76.19%
	1100	78.10%
	1300	87.62%
	1500	83.81%
120	500	69.52%
	700	80.95%
	900	66.67%
	1100	84.76%
	1300	85.71%
	1500	85.71%

3.3 the improvement

In this case, the method in paper for NN4---encoding is used to improve the accuracy for the model. Because of this data set is not so big, so we need not to delete too many lines of inputs to control the number of inputs. The most important factor is believed to improve the accuracy is the number of the feature and the method that they are inputted (Bustos & Gedeon, 1995). This table (Table 2) represents the different accuracy when controlling the number of inputs and we can find out that there are some improvements however they are not so obvious. The accuracy is nearly as same as the original method. So the conclusion is that encoding is more suitable when using the huge amount of data and the input is very complex. Small data set and simple input will not be affected deeply by this method, so that we can see there are still some gaps between the result and the result in the research report. The confusion matric shows that the two results are all predicted and both of them have good prediction. The bottom right and up left corner of numbers are the most so that it means the accuracy is good. And for this report, the genetic algorithm implies a new method to improve the parameters instead of trying those by hand. The algorithm will help us to find the most suitable parameters for the neural network and the LSTM are also help to deal with the huge amount of data. However, this data set have only about 400 columns so that the conclusion is not as good as expected. And the mission is classification so that the LSTM can not have the best influence. The LSTM algorithm is better at dealing with text.

Table 2. accuracy of improved feature

features	accuracy
20	81.90%
22	87.62%
24	80.00%
26	83.81%
28	87.62%
30	82.86%
32	85.71%
34	84.76%

3 Results and Discussion

3.1 results

From the table 1 it is obvious that the result of the neural network before improving is nearly 80%. In the testing it is observed that the result is floating from 79% to 85%. When the number of hidden neural is about 100 the result is the best. And the epoch is 1500 is the best condition to have the best accuracy. After improving, the result is nearly 90%. The most suitable number of feature is about 20. So that it can be concluded that the accuracy of final result is nearly 90%. It is not a very low percentage. However, comparing to the result of the research paper, 96%, there are still a distance. So it will be analyzed below. We can see clearly that the each step of the loss of the prediction. After comparing those loss, it will pass the best conclusion, which means the least of the lost, to the train the LSTM. And then the best parameters will be used to do the testing and we can see that the lowest loss is 0.0854. And when training the dataset, the loss is 0.0962.

3.2 analysis

The reason why this result is not as good as the research result can be guessed that the neural network is too simple. So more layers are tried to improve the accuracy. However, the result shows that the accuracy did not increase but decrease instead. So the second step is to change the number of hidden neural. And we can see that increasing the number of hidden neural properly will lead to the increasing of accuracy. And the increasing of epoch will also help with the test accuracy. The reason is that increasing the number of training and testing would help the accuracy. And then the improvement of input using the encoding method also help a little about the accuracy but not a very important factor because the data set is not big and complex enough. So it is not the most suitable method. Because of the complex degeree of the dataset, the original accuracy is not too low, but there are still a gap between this result and the research report. The result of the combination of these two algorithm is not as good as imagination because the data set is not big enough and the classification is not the most suitable problem for LSTM to solve. But this two algorithm do have some help and if we can use another bigger data set the conclusion will be much better than this one.

3.3 result comparison

It can be seen clearly that the accuracy of the classification task in the research is very high which can reach 96% (Kim & Park, 2003). However, our prediction accuracy is about 90%. The technique is not complex, but not so suitable for this dataset though this method is useful for classification questions. It is also obvious that the distance between the original and the research paper is not so huge. This is the size of this data set is small and this problem is not difficult to do. And we can realize that there are more method and improvements can be done on the setup of the neural network and the method to train the dataset to be get better accuracy and to suitable for other conditions. The two pictures below shows the result before improvement and after. In this report, when comparing the loss with the other conclusion, we can see that different parameters indeed have some differences in loss. But we can see that the differences are not obvious, because the dataset, as it said in previous, is not big enough to see the differences.

4 Conclusion and Future Work

4.1 conclusion

We can conclude that the encoding method that we learn from the paper is useful to get higher accuracy, but it is not the most suitable method in this case. Sometimes some little changes on the neural network, like the number of the hidden layers and the number of epoch will have a huge influence in the final accuracy, although it can not be guessed increase

or decrease in the first time. So a lot of training and testing is necessary. And also the suitable improvement method for the data set is important or the improvement may be very small or even decreasing.
In this report, the result shows that different data set all have their own suitable neural network and improve methods. Some of the neural network is for regression problem and some methods is for the huge data set with complex input. And the dataset which are without missing value is easy to do the preprocessing and all the data should be changed into number to get the easier classification. And also the lines of the data influence the speed of training and testing a lot. The split of training and testing is also an important factor which affect the accuracy.
In this report, we use two algorithms to improve the conclusion of the neural network. First one is the genetic algorithm and it is used to train the parameters and the second is LSTM and it can improve the data size that can be bigger. But the conclusion is not so obvious but we can improve that by using bigger dataset or using text data.

4.2 future work

It is obvious that we still have a lot work can be done. First, the result now is not the best result and there are still a lot of techniques can be used to improve the final result. Here are some directions that could be improve in the future. First, the setup of neural network can be improved. BP methods and many other methods can be used to improve it. The other factor including how to train the data and how to test the weight of the neural network. These are all aspects that can be considered to improve the final result of the classification. And also, the pytorch is not the only tool that we can use to set up the neural network, so other tools can be learned for improving the conclusion and give the more proper classification. In this report, we have know the reason that that the algorithm cannot be seen obvious improvement. So in the next step, we can use a bigger dataset or use this algorithm to deal with the problems related to text because LSTM are better to deal with those problems.

References

1. Bustos R.A.& Gedeon T.D. (1995) Decrypting Neural Network Data: A Gis Case Study. In: Artificial Neural Nets and Genetic Algorithms. Springer, Vienna
2.Kim, H. & Park, H.(2003).Data Reduction in Support Vector Machines by a Kernelized Ionic Interaction Model. Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining; Philadelphia : 507-511. Philadelphia: Society for Industrial and Applied Mathematics.