Application of Artificial Neural Network for Analysis of the Seismic Bumps Data Set

YUTING ZOU

Research School of Computer Science, Australia National University
u6342733@anu.edu.au

Abstract. This report shows result of application of neural network to predict the condition of the rock burst hazard, using the seismic-bumps dataset. This report compared its result with a relevant paper which using the same dataset and analysis the reason why the model in our paper performs worse. The bimodal distribution removal technique is used in order to improve the result. This paper explains why this method does not sufficiently improve the performance and propose the possible way to solve this problem in the future. This paper tries to use the genetic algorithm to do the features selection to decrease the training time and improve the neural network, and the result shows that this algorithm slightly improves the accuracy. The details are explained in this article.

Keywords: neural network, microseismic hazard prediction, classification, genetic algorithm

1 Introduction

1.1 About the Dataset

In this report, I choose the seismic-bumps dataset and try to use the neural network to solve a classification problem. Mining is a kind of activity by which could people get the coal, oil, and gas from nature. These resources are essential for the development of a country and people's everyday life. Safety has always been the most important issue in the mining industry, especially for underground mining. Although the modern mining industry is more secure than it used to be with the development of the technology, accidents often occur. Therefore, a practical and sufficient prediction system is necessary.

"The data set describe the problem of high energy (higher than 10^4 J) seismic bumps forecasting in a coal mine." [1] This dataset has eighteen input, which clearly recorded the situation of each mining activity. There are only one output values 0 or 1 to predict whether the dangers have great possibility to occur. The dataset has twenty-five hundred and eighty-four rows. The dataset clearly stated that "each row contains a summary statement about seismic activity in the rock mass within one shift (8 hours)" [1].

1.2 Related work

Many researchers use various kinds of methods to solve this problem. There is a paper uses the rule induction algorithm [2], in which compared the result of with the use and without the use the of rule filtration algorithm. This paper shows the possibility of solving the problem by using rule induction method. This paper clearly shows that the filtration algorithm could slightly increase the accuracy of the prediction as well. Another method using the induction and pruning of classification rules [3] to predict the microseismic hazards in coal mine. This paper compared the performance of the rule-based classifier with the classifiers using decision tree induction algorithm and neuro-fuzzy algorithm. The paper presents that this new method performs well than the prediction system at that time (2011). More recently, M.Sikora and B.Sikora introduce a new method which combine three techniques: regression rule conduction, KNN, and the time series forecasting by means of the ARIMA methodology together to do the analyzing. They use the M5 algorithm as a basic model as well [4].

1.3 Method brief introduction

This report tries to use the neural network to build an efficient predictive model. A two-layer neural network is built here as a classifier. Genetic algorithm is used to do the feature selection in order to improve the performance of the

classifier. I Override the initialization function along with the forward function to build a feed forward neural network. The loss function and the optimizer are defined after defining the neural.

2 Method

2.1 Data preparation before analysis

Before analyzing the data, the first step is to read the data from the file and transform it into the "dataframe" format. Since this file is a "arff" format, I use a function to read it and add the columns to make it a standard dataframe. Converting the data from other data type to the float type is needed with the requirement of standardization process.

When we do the data analysis in machine learning, normalizing the data is an essential step since different evaluation method often have different dimensions and dimension units. This kind of situation would exert influence in the process of data analyzing and may affect the final result. Since the unit of the data is not the same, the range of some data may be particularly large, resulting in a slow neural network convergence and long training time. In this report, some proper attributions are chosen to do the minimal-maximum normalization.

2.2 Oversampling method

When I preprocess the dataset, I find that the 1 class is a minority class. The dataset has 2584 rows of data but there are only 170 data belong to this class. In order to let the model learn the minority class better, since the dataset is not quite large, I use the over sampling method to increase the number of minority class [5]. I make all these 170 data repeatedly appeared in the dataset, at this time, there are around 510 data which has the 1 value in attribute "class", and the total number of the dataset becomes 2924.

2.3 Genetic algorithm

Genetic algorithm (GA) is metaheuristic inspired by the process of natural selection [6]. Genetic Algorithm (GA) originated from computer simulations of biological systems. It is a stochastic global search and optimization method that has evolved from the biological evolution mechanism of nature [6].

There are two ways to combine the neural network with the genetic algorithm [7]. Firstly, we could use the genetic algorithm to preprocess the data, and then use the neural network to solve the problem. The second one is cooperation. Under the structure of the neural network, we use the genetic algorithm to determine the link weights or optimize the neural structure, then use the back-propagation algorithm to train the neural network. In this paper, I use the genetic algorithm to do the features selection, try to reduce the training time by finding out the features which could greatly influence the target.

The GA algorithm starts from representation [8], selecting a subset of the stress features. Secondly initialize the population. And then calculate the fitness and evaluate the fitness value for each chromosome in the population. If the condition is satisfied, we would terminate. If the condition is not satisfied, we need to apply crossover and mutation to the population. Then we evaluate the fitness value for each new chromosome in the population and finally select the chromosomes that well proceed to next generation. And we justify whether the condition is satisfied again. If the condition is satisfied, we terminate the evolution, if not, we continue the crossover, mutation, evaluate fitness value, and the select process until we meet the requirement.

In this paper, we use neural network as the classifier. The training accuracy here is used as the fitness value to decide which features should be selected. By selecting the group of features which have the highest fitness value, we could find a combination of attributes that gives the highest accuracy.

2.4 Construct neural network

Some variables used in the neural network should be defined before building the neural network. As it is mentioned above, there are eighteen input neurons and two output neurons. The learning initially defined as 0.001. Since if the

learning rate is too large, then loss will burst, if the learning rate is too small, the waiting time is particularly long. We found that when the number of training reaches 1000, the loss will drop to a relatively small value. So, I defined the number of training as one thousand.

Then construct the neural network class and linear hidden layer output [9]. After that, constructing the "forward" function by defining the input of hidden layer, the activation function and the output of the output layer. The module would call the "forward" function to do the forward propagation and build the neural network. The "CrossEntropyLoss" function is used as a loss function. Since the dataset is not quite large, I use full batch training in the training process, so I use Rprop to individually update individual weights based on gradient symbols and targeted.

2.5 Train and test the neural network

To get the training set and testing set, firstly all the data is shuffled [9], then the dataset is randomly split to eighty percent for the training set and twenty percent for the testing set. Since the dataset is not quite large as it is mentioned above, in order to learn the model better, we could not set the size of the training set too small. We could not set the testing set too small as well in order to reduce the chance in prediction. I balance the two sides and finally eighty percent data is used as a training set, twenty percent data is used as a testing set. The dataset has nineteen attributes, the previous eighteen attributes are inputs and the last one is output which has two values. According to this, I split the training data into input and target, which contains eighteen columns and one column separately. The last step is creating tensors to hold the inputs and output and wrap them in variables. And this is the end of the preprocess.

In the training process, I use full batch learning method, and trained all the training set a thousand times since the dataset is not quite large. Each time, I get the loss and count the accuracy according to the predicted result and the actual data. Then the back-propagation algorithm is used and the step function on an optimizer is called to update the weights to make the error in the proper range.

The test set is a set of data which is used to evaluate the performance of the classifier. And the test set is used as the input of the neural network, after the processing of the neural network, we get the predicted result and then compare it with the actual result and finally get the testing accuracy. A confusion matrix is built for further analysis.

2.6 Bimodal distribution removal

In this paper, I try to use the bimodal distribution method [10] to improve my classifier model. This method tries to do some cleaning to the dataset to remove the noise and outliers which would decrease the accuracy since it is hard to predict. They calculate the mean of the error (here, the error is loss) and remove those data whose loss is larger than the average loss from the dataset. In this way, the training set is reduced without the decreasing of the accuracy. Here is the step of this algorithm [10] [11]:

1. Train the whole dataset.
2. Wait until the variance of the errors over the training set vts is below 0.1, which indicates the two error peaks have formed.
3. Calculate the mean error ts over the training set.
4. The pattern in the high error peak are outlier candidates. All those patterns with error greater than $\bar{\delta}_{ts}$ are taken from the training set.
5. Calculate the mean $\bar{\delta}_{ss}$, and the standard deviation σ_{ss} of this subset.
6. Permanently remove all patterns from the training set with $\text{error} \geq \bar{\delta}_{ss} + \alpha\sigma_{ss}$ where $0 \leq \alpha \leq 1$.
7. Repeat the steps 2-6 every 50 epochs, if vts <= 0.01, then the training is halt.

2.7 Evaluation indictors

The first indicator I use is accuracy. The testing accuracy is about the number of predicted correct elements divides the total number of the test set.

The dataset we use in this paper is imbalanced, and classifiers can achieve high accuracy just by classifying all classes into majority classes. However, classifiers at this time are not meaningful. At this point, we need Precision, Recall, harmonic F, and ROC [13]to measure the effectiveness of the classifier. Precision is a measure of accuracy, indicating the proportion of actual positive examples in positive examples [12]. Recall rate is a measure of coverage. It indicates the number of positive cases divided by the number of results that should be returned and measures the classifier's ability to identify positive examples.

F1 score is a harmonic mean of the precision and recall [12]. We hope both the precision and recall could be closer to 1, however, these two values are contradictive sometimes, so f1-score is necessarily used since it is a combination of these two values. When F1 is higher, it shows that the experimental method is more ideal. The sklearn is used to get the precision, recall and f1 value.

The last indicators I use is the AUC value, which is the area under the receiver operating characteristic curve (ROC) [13]. It is a probability compare to the random probability. AUC can avoid converting predicted probabilities into categories, which could decrease the influence of the threshold.

3 Result analysis and discussion

3.1 Compare the results with or without the use of GA

In the first part, I would like to show how the genetic algorithm improve the neural network.

Here is the testing accuracy and the AUC value.

Chart 1

	Number of features	Testing accuracy	AUC value
With the use of GA	10	88.24%	0.676
Without the use of GA	18	85.57%	0.639

Chart 2

The classification report without the use of genetic algorithm			
	precision	recall	F1-score
0	0.87	0.94	0.90
1	0.56	0.34	0.42

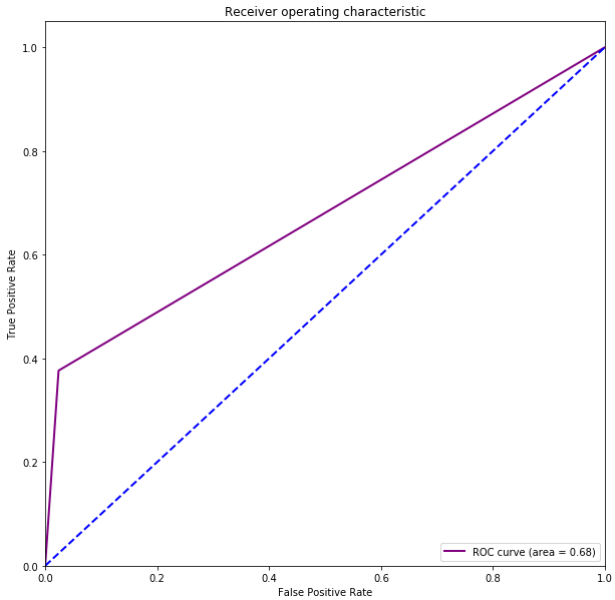
Chart 3

The classification report with the use of genetic algorithm			
	precision	recall	F1-score
0	0.87	0.96	0.91
1	0.65	0.37	0.47

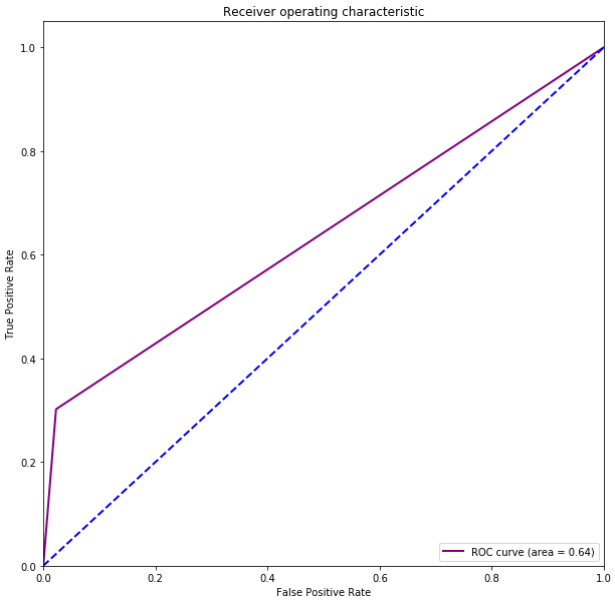
Chart 1 clearly shows that the neural network classifier has a good prediction accuracy, which is more than 80%.
From this chart we could say that the genetic algorithm slightly increases the testing accuracy and the AUC value.

Chart 2 and chart 3 above shows the classification report of the model use different features. We could find that the f1-score value in chart 3 is larger than the f1-score in chart 2, which indicates that genetic algorithm slightly improves the performance of the model.

Graph 1 below shows the ROC with the use of genetic algorithm, graph 2 below shows the ROC without the use of the genetic algorithm. We could find that although the number of features decrease, the AUC value increase, which indicates the model has a better performance and features selection is meaningful.



Graph 1



Graph

3.2 Compare the result with and without the use of the BDR algorithm

In this part, I would like to compare the performance of the model with and without the use of the BDR algorithm. Both of the two models in this part use the genetic algorithm to do the features selection.

The indicators without the use of BDR algorithm

	Testing Accuracy	Precision	Recall	F1-Score	AUC
0	88.24%	0.87	0.96	0.91	0.679
1		0.65	0.37	0.47	

The indicators with the use of BDR algorithm

	Accuracy	Precision	Recall	F1-Score	AUC
0	85.93%	0.87	0.97	0.92	0.635
1		0.68	0.30	0.42	

(The number of training times is 1000)

From the chart above, we could clearly find the BDR algorithm did not sufficiently improve the performance of the classifier. On the contrary, the implement of this algorithm slightly decreases the accuracy in some extent since both the AUC value and the testing accuracy declines.

The reason is that not all the data being removed from the training set is noise. Under this kind of situation, the loss of the dataset is not conductive to learn all the patterns and build an efficient classifier model, which would lead to the low accuracy when testing the test set.

4 Comparative the results with the relevant paper

The paper I choose is “APPLICATION OF RULE INDUCTION ALGORITHMS FOR ANALYSIS OF DATA COLLECTED BY SEISMIC HAZARD MONITORING SYSTEMS IN COAL MINES”. This paper does the prediction in hourly horizon and shift horizon, changing many factors such as train set, test set, use or not use the rule filtration algorithm to do the comparation and finally get the results in different conditions. Here is the best result they get:

Classification results – test set – shift prediction horizon

Dataset (shift aggregation)	Accuracy (%) – „hazardous" state	Accuracy (%) – „non-hazardous" state
The results obtained without the use of rule filtration algorithm		
SC503	73.3± 0.00	93.6 ± 0.00
SC508-max	85.3 ± 0.01	71.9 ± 0.00
SC508-avg	70.1 ± 0.01	77.4 ± 0.01
The results obtained with the use of rule filtration algorithm		
SC503	74.4 ± 0.00	94.6 ± 0.00
SC508-max	84.6 ± 0.02	73.7 ± 0.00
SC508-avg	68.3 ± 0.05	79.5 ± 0.00

(This chart is on the P14 of the paper and called "table1", SC508 - 864 records, 97 of which were assigned to "hazardous" state, SC503 - 1097 records, 188 of which were assigned to "hazardous" state. [2])

This chart clearly shows that different training set and testing set would exert different results. From this paper, the highest accuracy of predicting the "hazardous" state is using the SC508-max dataset and rule filtration algorithm, which is around 84.6. The highest accuracy of predicting the "non-hazardous" state is using the SC503 dataset and rule filtration algorithm, which is around 94.6%.

In my program, genetic algorithm is used to do the features selection. And the precision of the "hazardous" state is around 0.65, and the precision the "non-hazardous" state is around 0.87. Under both of the two conditions, my classifier performs worse.

The first reason makes my classifier performs worse than this paper is the different training set and testing set. From the chart shows in the relevant paper, we could find that different data partitioning may exert difference in the final accuracy. Different proportion that the 'hazardous' state takes would also make difference in the result.

The second reason is that this paper uses the k-fold cross validation. K-fold cross validation could be used to test the trained model, mainly for testing supervised learning modelling results, along with the degree of deviation of the model parameters. It could adjust the parameters of the classifier and get the practical group to make the classifier performs better.

5 Conclusion and future work

This article presents the results of using the neural network as a classifier to predict seismic hazard in the mining activities. The bimodal distribution removal algorithm is implemented in this program to find out if it could improve the performance of the classifier. This article uses several indicators and tries to figure out the behavior of the classifier from different aspects.

This report firstly does the analysis on the original dataset and find that the dataset is imbalanced. In order to let the model learn the minority class better, the over-sampling method is used, and the copies of the minority class data are added to the original dataset.

This article then tries to use the genetic algorithm to do the features selection. From the result we could find that this algorithm slightly improves the testing accuracy of the dataset, from 85.57% to 88.24%. and the AUC value is also slightly improved, from 0.64 to 0.68. However, we could also find that both the testing accuracy and the AUC value are not improved quite much. Since there are not too many attributes in the dataset, only 18 inputs and 1 output, the advantages of the genetic algorithm may not be obvious. But from this paper we could see that the genetic algorithm has the potential ability to improve classifier accuracy by selecting attributes which could greatly influence the target.

The article then compared the model which use the genetic algorithm with the relevant paper. And find that the classifier present in the relevant paper performs better, whose best accuracy is more than 90% when predicting the non-hazardous state and more than 70% when predicting the hazardous state with the use of SC503 dataset [2].

The BDR algorithm is implemented in order to improve the performance of the classifier. However, the result did not reach our expectation. This algorithm slightly decreases the accuracy of our classifier but not too much.

There are lots of things we could do to improve our classifier in the future. Firstly, we could use more sufficient sampling method to improve the structure of the dataset [14]. We could attempt to generate artificial data sample with the use of the Synthetic Minority Over-sampling Technique, which is known as SMOTE. The SMOTE is an over-sampling method. It constructs new, small class samples instead of generating copies of existing samples in the minority class. The data constructed by this algorithm is not real exists in the original dataset. This would decrease the overfitting

problem [14]. Secondly, a more efficient way of reducing the size of the training set could be used aiming to make the classifier behaves better. We need to use a more practical way to find the noise then remove it from the dataset without decreasing the prediction accuracy. Thirdly, we could not only use the genetic algorithm to do the features selection, we could also use it to find the most suitable parameters of the neural network. By adjusting the parameters in the neural network, this classifier model could have a better performance.

6. Reference

[1] Archive.ics.uci.edu. (2018). UCI Machine Learning Repository: seismic-bumps Data Set. [online] Available at: http://archive.ics.uci.edu/ml/datasets/seismic-bumps [Accessed 29 Apr. 2018].

[2] Sikora M, Ł. Wróbel. Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines[J]. Archives of Mining Sciences, 2010, 55(1):91-114.

[3] Sikora M. Induction and pruning of classification rules for prediction of microseismic hazards in coal mines[M]. Pergamon Press, Inc. 2011.

[4] Sikora M, Sikora B. Improving prediction models applied in systems monitoring natural hazards and machinery[J]. International Journal of Applied Mathematics & Computer Science, 2012, 22(2):477-491.

[5] Brownlee, J. (2015). 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ [Accessed 29 Apr. 2018].

[6] En.wikipedia.org. (2018). Genetic algorithm. [online] Available at: https://en.wikipedia.org/wiki/Genetic_algorithm [Accessed 28 May 2018].

[7] Blog.csdn.net. (2016). Neural network with genetic algorithm – CSDN Blog in China. [online] Available at: https://blog.csdn.net/u011001084/article/details/49335201 [Accessed 28 May 2018].

[8] Wattlecourses.anu.edu.au. (2018). Wattle: Log in to the site. [online] Available at: https://wattlecourses.anu.edu.au/pluginfile.php/1762198/mod_resource/content/0/IN1-%20GAs%20for%20Feature%20Selection.pdf [Accessed 28 May 2018].

[9] Wattle. (2018). lab2_Answer.zip. [online] Available at: https://wattlecourses.anu.edu.au/course/view.php?id=22184 [Accessed 29 Apr. 2018].

[10] Slade P, Gedeon T D. Bimodal distribution removal[J]. Lecture Notes in Computer Science, 1993:249-254.

[11] Wattlecourses.anu.edu.au. (2018). Wattle: Log in to the site. [online] Available at: https://wattlecourses.anu.edu.au/pluginfile.php/1664419/mod_resource/content/2/NN8_finit.pdf [Accessed 28 May 2018].

[12] En.wikipedia.org. (2018). Precision and recall. [online] Available at: https://en.wikipedia.org/wiki/Precision_and_recall#Precision [Accessed 29 Apr. 2018].

[13] En.wikipedia.org. (2018). Roc. [online] Available at: https://en.wikipedia.org/wiki/Roc [Accessed 29 Apr. 2018].

[14] En.wikipedia.org. (2018). Oversampling and undersampling in data analysis. [online] Available at: https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis [Accessed 29 Apr. 2018].