Using Bimodal Distribution Removal Method to Update Neural Networks

Zhuoxun Zhao
Research School of Computer Science
Australian National University
u6281707@anu.edu.au

Abstract. The aim of this project is to use the Bimodal Distribution Removal as the outlier removal method to improve the quality of simple feedforward neural network. Genetic Algorithm will be used to reduce the number of redundant features. The given dataset which trained in these neural networks is acquired from UCI Machine Learning Repository. It is labelled as the "Mushroom Data Set". Two networks include in this paper, one is three-layer feedforward Multilayer Perceptron, another one is the former neural network but use Bimodal Distribution Removal approach to remove the noisy data. Compare these two networks with the Naive-Bayes and General Bayesian network and discuss the result. The outcome of the final model use BDR method to reduce the size of dataset and Genetic Algorithm to select the minimal feature subset – approximately 97% accuracy in the testing dataset

Keywords: Feedforward Neural Network, Bimodal Distribution Removal, Naive-Bayes, Genetic Algorithm, Feature Selection

1. Introduction

The dataset used in this report labelled "Mushroom Data Set", acquired from UCI Machine Learning Repository [1]. This dataset contains the records of 8124 mushroom from the Audubon Society Field Guide to North American Mushrooms. The dataset recorded in 1981, which is comprised of 22 hypothetical samples and a binary prediction to describe the mushroom is poisonous or edible. Dataset was picked as it contains less missing values and each pattern is identified as definitely edible or definitely poisonous. The labels and features of the dataset are clearly, only one attributes include the missing value and there are no species was described as unknown edible.

There is no simple rule for determining the edibility of a mushroom, the mushroom dataset contains 22 features and uses these to predict the edibility. The edibility of a mushroom is either poisonous or non-poisonous, and this report uses the neural network to predict the result. This study uses Genetic Algorithms (GA) as a tool for feature selection [2]. Feature selection aims to reduce the number of redundant features without losing performance and accuracy. There are two main neural network structures tested in this report, consist of a simple Multilayer Perceptron and a Multilayer Perceptron with bimodal distribution removal method [3]. The simple Multilayer Perceptron used three-layer feedforward structure and included an input layer, a hidden layer and an output layer. The structure of the second neural network is similar to the first one but use a cleaning up method to detect the noisy data from the original dataset. The critical difference between two neural networks is the algorithm will remove the incorrect patterns before it is fed into the network. Rather than train the original data, the dataset with few erroneous patterns could speed up the learning rate of majority data and reduce the effect of overfitting [3]. This report will discuss the detailed

cleaning method in "Method" section, analyse the test result and compare with the benchmark method in "Results and Discussion" section.

1.1 Dataset description

The dataset used here is comprised of 8124 instances each of which has 22 features as listed below, the missing value only occurred in the No.11 attribute and was denoted by the symbol "?" [1].

ATTRIBUTE	CLASSES
cap-shape	6
cap-surface	4
cap-color	10
bruises	2
odor	9
gill-attachment:	4
gill-spacing:	3
gill-size:	2
gill-color:	12
stalk-shape:	2
stalk-root:	7
stalk-surface-above-ring:	4
stalk-surface-below-ring:	4
stalk-color-above-ring:	9
stalk-color-below-ring:	9
veil-type:	2
veil-color:	4
ring-number:	2
ring-type:	8
spore-print-color:	9
population:	6
habitat:	7

2. Method

In this section will discuss the normalization method of the dataset, the details of the feature selection method and the evaluation method. Moreover, introducing the details of the neural network structure, the cleaning up approach and the benchmark algorithm. Genetic Algorithm will be used to reduce the number of redundant features.

2.1 Pre-processing method

The original type of attributes in "Mushroom Data Set" is the alphanumeric character, it uses a letter of the alphabet to describe the attribute Information. For example, the type of mushroom veil contains two types, this dataset use "p" show the meaning of "partial", and the initial "u" to replace the word "universal". The pre-processing method used here is to transform non-numerical labels to numerical labels, it is similar to the "LabelEncoder" approach in the Scikit-learn library. The encode theory has normalized the labels by encoding the value of labels between 0 and the number of classes this label included, such as turn [brown, orange, white, yellow] into [0, 1, 2, 3].

2.2 Genetic Algorithm and Feature selection

Genetic Algorithm is a search method based on the paradigm of natural selection and population genetics. It uses a fixed length binary string to represent a possible solution for a problem domain. There are three main operators applied in Genetic Algorithm, selection, crossover and mutation. Each possible solution or individual could be evaluated by a fitness function. Selection operation select the individuals based on its fitness values, high-fitness individuals do tend to be picked as the parents. In crossover step, a pair of selected parents will exchange their information and generate the new individuals. Mutation operator involves a probability that a bit of selected individual will be changed from its original state. The formula of fitness function is [2]:

$$\text{fitness(x)} = \text{acc(x)} \times \rho_{x,y}$$

Where acc(x) is the classification performance on training set using individual x. There are N total number of patterns, C is the number of correct classified patterns.

$$\text{acc(x)} = \frac{C(x)}{N}$$

$\rho_{x,y}$ is a correlation coefficient between feature x and label y. It is used to measure a relationship between label y and a feature subset x. If the value of $\rho_{x,y}$ is close to one, it means has a strong linear association between the feature x and the label. If $\rho_{x,y}$ equals zero, there is no linear association between the dependent variable y and the independent variable x.

Feature selection is an important part of the classification problem, the choice of feature subset could affect the accuracy, training time and the running cost [2]. The aim of feature selection is to select a minimal number of features to represent a data to be able to distinguish from other classes.

In this paper, use a chromosome to represent the feature subset. If the indicate F_x is 0, means this feature is not in the subset. If the indicate F_x is 1, means this feature is in the subset. F_x is the x_{th} feature of the chromosome F.

2.3 Simple feedforward neural network

The structure of the simple neural network is a three-layer feedforward neural network, the input layer size is 22 as there are 22 attributes, the size of hidden layer is 200, and the output layer size is 2 because the mushroom is either poisonous or edible. Use rectified linear units (ReLUs) for the hidden layers and select stochastic gradient descent method as the optimizer.

2.4 Bimodal distribution removal method

Bimodal Distribution Removal (BDR) is a method to clean up the noisy data and improve the efficiency and accuracy of the neural network. It uses outlier detection method to evaluate the quality of the dataset, which train the neural network and use the training dataset to measure the prediction [3]. Unlike the formal training, BDR could seem as a pre-training, it only produces about 50 epochs to train the neural network. The algorithm will remove the sub datasets with the high error and repeat until the variance of the network below a constant. BDR use an equation to distinguish the" good" subset and the "bad" subset:

$$\text{error} = \overline{\delta}_{ss} + \alpha\sigma_{ss}$$

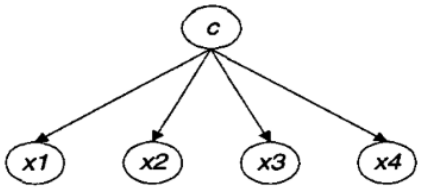
In this equation, $\overline{\delta}_{ss}$ is the mean error of all the sub datasets, α is a coefficient to control the remove speed, and σ_{ss} is the standard deviation. The range value allowed for α is zero to one.
Standard deviation is a measure in statistic area that used to quality the quality the amount of variation of datasets. The formula of the sample standard deviation is:

$$\sigma_{ss} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1}}$$

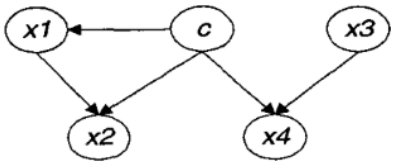
It is used to describe how the data points tend to the mean value of the datasets. In Bimodal Distribution Removal (BDR) method, the standard deviation is a supplementary parameter to decide how many patterns need to remove from the original dataset.

2.5 Benchmark method

The paper used to benchmark this project is labelled, "Comparing Bayesian Network Classifiers". This paper introduces and compares the Naive-Bayes classifier and its extension methods. Select two most representative approaches from the paper as the benchmark. There are Naive-Bayes and General Bayesian network. Naïve-Bayes is one of the most famous supervised machine learning methods, it maintains a simple structure that the label stored in the parent node and connect all the leaf nodes. The leaf nodes contain the information of attributes. There is no connection between any two leaf nodes. Below is a simple example of Naive-Bayes:



General Bayesian network is a probabilistic graphical model that use a directed acyclic graph to show the conditional dependencies between different attributes. It treats equality to all the classification nodes and uses Markov blanket to represent the connection of these nodes [4]. A simple General Bayesian network looks like:



Where the node c is the label node, and the $x_{1\ldots}x_4$ represent the attribute nodes.

2.6 Evaluation method

The evaluation method used in the benchmark report is prediction accuracy, is check the error accuracy in the test dataset. Error accuracy is a fairly straightforward approach that shows the quality and reliability of the trained neural network, it fed the test dataset in the network, and output the predict results. The ratio of predict result to actual labels is the prediction accuracy.

$$accuracy = \frac{correctly\ classified\ patterns}{total\ patterns}$$

Another evaluation method is F1-measure, which is a harmonic mean of precision and recall. Precision is the fraction of retrieved patterns that are relevant. In a binary classification problem, is the number of correctly classified "1" patterns divided by the number of patterns labelled by the system as "1". Recall is the fraction of relevant patterns that are retrieved. In a binary classification problem, is the number of correctly classified "1" patterns divided by the number of "1" patterns in the dataset [5].

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3. Results and Discussion

In this section, will analyze the result of the simple feedforward neural network and compared it with the neural network with the Bimodal distribution removal method. Compare the result of the standard pre-processing method and feature selection approach. Contrast the benchmark algorithm with the neural network, and the discuss the reasons behind these.

There are 8125 data rows in the Mushroom Dataset. In this section, randomly split the dataset into the training set (80%) and testing set (20%).

3.1 Simple feedforward neural network

The following figure shows the change of accuracy with training times in the training phase. The training phase will run 700 epochs to improve the neural network, output the accuracy of the training dataset after every 100 runs.

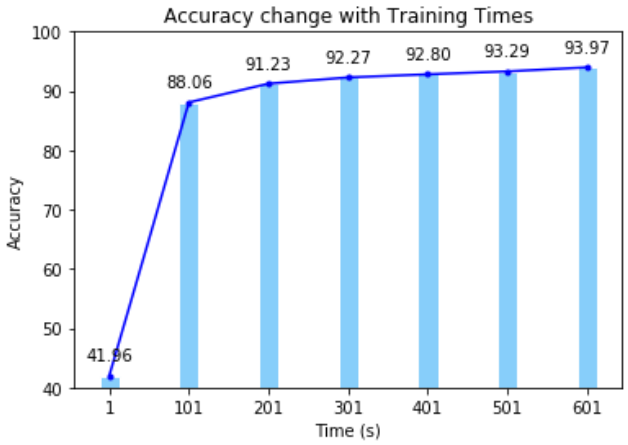


Figure 1: Simple Neural Network

We can find the accuracy of the neural network increase very quickly at the first 100 epochs, and the growth rate slowed in the last times. The loss rate of the neural network reduced to 0.18 after 700 runs. Use the same preprocessing method and network to predict the testing dataset. The final accuracy of the testing dataset is 94.87%, and the F1 score is 0.948.

3.2 Neural network with bimodal distribution removal method

The mark equation of the BDR method is:

$$\text{error} = \overline{\delta}_{ss} + \alpha\sigma_{ss}$$

The range value allowed for α is zero to one. The value of α is used to control the removing speed, and there was a negative correlation between the number of delete patterns and this coefficient value. With the decrease of value α, the delete standard line decline, thus more patterns were removed from the training dataset. First, initial the value of α to one, and the following figure shows the result of the training phase.

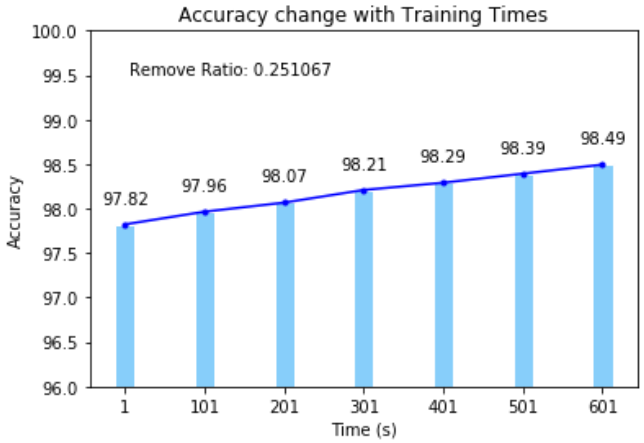


Figure 2: Neural Network with BDR method

The definition of Remove Ratio is $1 - \frac{Number\ of\ New\ TrainingDataset}{Number\ of\ Old\ Training\ Dataset}$, it is used to describe the percentage of patterns removed from the training dataset. The parameters used in this network is same as the former one. Compare with the simple feedforward neural network, could find the initial accuracy of the neural network is 97.55%, which is much higher than the formal network. The final accuracy of training dataset is 98.27%, also higher than the simple feedforward neural network. The loss rate is reduced to 0.06, the accuracy of the testing dataset is 95.97%, and the F1 score of the testing dataset is 0.965.

Change the value of α to increase the Remove Ratio, that delete more patterns. Result shows in the following table, the BDR_n in the table means use the bimodal distribution removal method to remove the noisy data, and the number n is the value of α selected.

Name	**Dataset Size**	**Train Accuracy**	**Train Loss**	**Test Accuracy**	**Test F1 Score**
Original	6562	93.97%	0.18	94.87%	0.948
BDR_1	4913	98.49%	0.06	95.97%	0.965
$BDR_{0.7}$	3265	99.57%	0.02	94.02%	0.959
$BDR_{0.5}$	2833	99.96%	0.01	93.84%	0.948
$BDR_{0.3}$	3137	99.87%	0.01	93.78%	0.948
$BDR_{0.1}$	2752	100%	0.01	93.53%	0.947

Table 1: Accuracy of Different BDR Parameter

According to the above table (Table 1), we could find the training dataset size is reduced with increasing of the value of coefficient α. The accuracy of the training dataset increased, but the

accuracy of the testing dataset and the F1 score of testing dataset reduced. One available reason is the dataset is too small, that dramatically increases the overfitting effect [3].

3.3 Neural network with Feature selection

The follow figure shows the change of accuracy with training times in the training phase, the parameters and training epochs are same with the section 3.1.

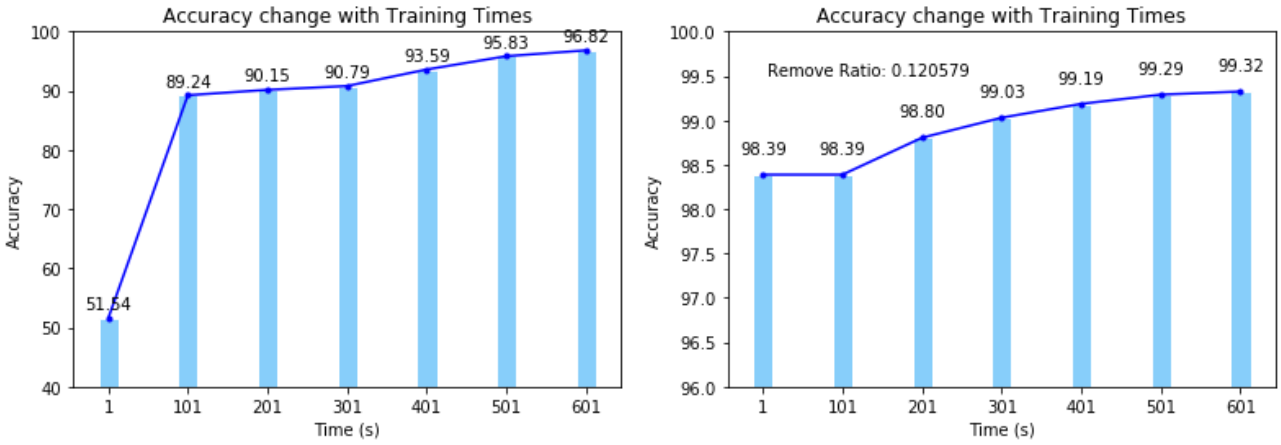


Figure 3: Neural Network with Feature Selection

The left chart in Figure 3 is used Genetic Algorithm to reduce the features and then run the models in section 3.1. The result is similar to the simple neural network, the accuracy of the training dataset increases very quickly at the first 100 epochs, and the growth rate slowed in the last times. The accuracy of the testing dataset is 96.85%, and the F1 score is 0.971.

The right chart in Figure 3 is used Genetic Algorithm to reduce the features and then run the models in section 3.2. There is a significant reduction of Remove Ratio, from 0.2 to 0.1. One probable reason is remove the irrelevant features could reduce the noise of the dataset. The accuracy of the testing dataset is 96.89%, and the F1 score is 0.969.

Name	**Train Accuracy**	**Test Accuracy**	**Test F1 Score**
Simple Neural Network (SNN)	93.97%	94.87%	0.948
SNN with Feature Selection	96.82%	96.85%	0.971
SNN with BDR method	98.49%	95.97%	0.965
SNN with BDR method & Feature Selection	99.32%	96.89%	0.969

Table 2: Comparison between Different Models

According to the Table 2, both the Feature Selection and the Bimodal Distribution Removal method could increase the Accuracy and F1 score of the testing dataset. SNN with BDR method performs better than SNN with Feature Selection on the "Train Accuracy" part, but bad at the testing dataset. One probable reason is the BDR method reduce the size of training dataset, which increases the overfitting effect and affect the generalization of neural network [3].

3.4 Comparison with the benchmark paper

The General Bayesian network (GBN) method introduced in the benchmark, use the feature selection approach to choose only 5 of the 22 features. The author of "Comparing Bayesian Network Classifiers" try different threshold settings and find the above parameters is the best one [4]. In this paper, the author only uses the Accuracy of testing dataset to evaluate the models.

Method Name	**Accuracy**
Simple neural network (SNN)	94.87%
SNN with Feature Selection	96.85%
SNN with BDR method	95.97%
SNN with BDR method & Feature Selection	96.89%
Naive-Bayes	95.79%
General Bayesian network	99.30%

GBN performance best, the probable reason is the GBN produce outstanding results when the datasets are large, and it has used the dimension reduction method to remove the weak relevant features [4]. Compare with the above table, BDR method can increase the accuracy of the testing dataset, the result of simple feedforward neural network is worse than the Naïve-Bayes. After using the bimodal distribution removal method to except the noisy data, the accuracy of testing dataset growth to 95.97%, which is better than Naïve-Bayes. The outcome of the final model uses BDR method to reduce the size of dataset and Genetic Algorithm to select the minimal feature subset – approximately 97% accuracy in the testing dataset, the benchmark paper uses Naïve-Bayes method to achieve an accuracy of 95.8%.

4. Conclusion and Future Work

Having explored the dataset and train the neural network with and without the bimodal distribution removal method, we can confidently conclude that the BDR method can improve the accuracy of training and testing datasets. The experimental results show that use Genetic Algorithm to remove redundant features can perform satisfactorily in testing dataset. The advantages provided by the feature selection in this paper is use correlation with classification performance to measure the relationship between any feature and labels. The combination of feature selection and bimodal distribution removal method has allowed us to improve the quality of the neural networks. The highest accuracy of the classification achieved in this report is 96.89%, which is produced by the bimodal distribution removal and Genetic Algorithm.

For the future work, plan to find the connection between the best network accuracy and the optimal coefficient value α, and the optimal constant to stop using BDR method. Another area that could be explored is trying to remove the weak relevant features from the network, could pre-training the network and find the features with the highest error. Feature selection which was introduced in this paper needs to be explored further as the Mushroom dataset only has two classes. A future extension on the Genetic Algorithm part is solving the feature selection problem in aspects of abilities to handle multiple classes datasets.

References

[1] J. Schlimmer, "Mushroom Data Set," UCI Machine Learning Repository, 27 04 1987. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Mushroom.

[2] C. Nipadan and Y. Qi, "Genetic algorithms in feature selection," *Systems, Man, and Cybernetics,* vol. 5, pp. 538-540, 1999.

[3] P. Slade and T. Gedeon, "Bimodal distribution removal," *International Workshop on Artificial Neural Networks,* pp. 249-254, June 1993.

[4] J. Cheng and R. Greiner, "Comparing Bayesian Network Classifiers," *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence,* pp. 101-108, 1999.

[5] C. D. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval, New York: Cambridge University Press, 2008.