Classification of Heart Disease Using Neural Networks

Wei Yu (2731**), Hanxiao Liu (2731****)**

Submission date: 2016-05-27

Abstract

In hospital, it is essential and important to find a valid diagnosis method of heart disease to take preventive measures and make patients get treatments in time. To solve this problem, we try to build a classification neural network model by conducting a series of experiments. First, we select GRNN architecture as our neural network architecture by comparing the accuracy rate of each candidate architecture in five experiments. Second, based on GRNN model, three experiments with different test set rate are conducted and the test set rate of 20 percent are chosen due to the higher accuracy rate than other two competitors. Besides, 100 neurons are selected in the hidden layer of GRNN architectures after 12 experiments are performed through comparing the accuracy rate and R squared value. During the selection of parameters of neural network architecture, the accuracy rate of test set is improved step and step and the final accuracy rate is 0.649123 after determining neurons of the hidden layer through 12 experiments. Finally, we point out current limitations of the process of constructing a classification neural network model and propose future solutions.

Keywords: classification neural network, GRNN architecture, accuracy rate

1 Introduction

In current society, the incidence of heart disease is getting increasingly higher. Heart disease has become a terrible disease threating human's life. In UK and Austria, heart disease has become the top three causes of death [1]. The cause of heart disease is complicated. Many factors will contribute to heart disease. Therefore, it is essential to train a model to distinguish heart disease and the level of heart disease. In the process of our research, we have reviewed several articles about classification or prediction of heart disease using the data from Cleveland database. Ming et al. [3] deploy improved BP algorithm to do an intelligent diagnosis coronary heart disease. Ganesan [2] put forward a classification framework based on C4.5 algorithm for medicinal data and this framework propose a missing value replacement technique which can solve missing values in data sets.

In this paper, we adopt NeuroShell 2 as our experiment environment. Training samples are retrieved from UC Irvine Machine Learning Repository. 303 records of heart diseases are extracted. After preprocessing primary data. We eliminate the invalid data and 287 records are remaining. We therefore take aforesaid 287 records as our training data. To evaluate the classification result, the result set are exported into Excel 2013. It is assumed to be correct that if the absolute value of deviation from real data is less than 0.5. Accuracy rate of test result also can be further calculated. After retrieving training data and constructing evaluation standard, the neural network architecture, test set rate and neurons are also selected step by step through experiments. The concrete analysis process of selection is described in Section 3 and Section 4.

2 Data Sets

During our experiments, the data about heart disease is from UCI machine learning repository heart disease data set. There are four databases in heart disease data set. The four databases are Cleveland, Hungary, Switzerland, and the VA Long Beach. We use data from database Cleveland. This database contains 76 attributes, but most experiments prefers to use only 14 attributes in these attributes. The reason why we choose the data from database Cleveland is that the Cleveland database is the only that has been used by ML researchers. Availability of voluminous data in the medical domain has provided an opportunity of extracting useful information from that data [2]. There are 303 instances of heart disease in Cleveland database. The data of the 303 instances are preprocessed by the provider of the data. But there are some missing values in the 303 instances, which will make the instance invalid. Through removing instances with missing values, 287 instances are left, which will be regarded as experimental data in this paper. There are 13 inputs and 1 output in the experimental data. Each of output has 5 states to represent the condition of heart disease: 0, 1, 2, 3 and 4. The condition of 0 represents no heart disease, 1, 2, 3, and 4 represents the increased level of heart diseases respectively.

3 Training Issues

3.1 The selection of architecture

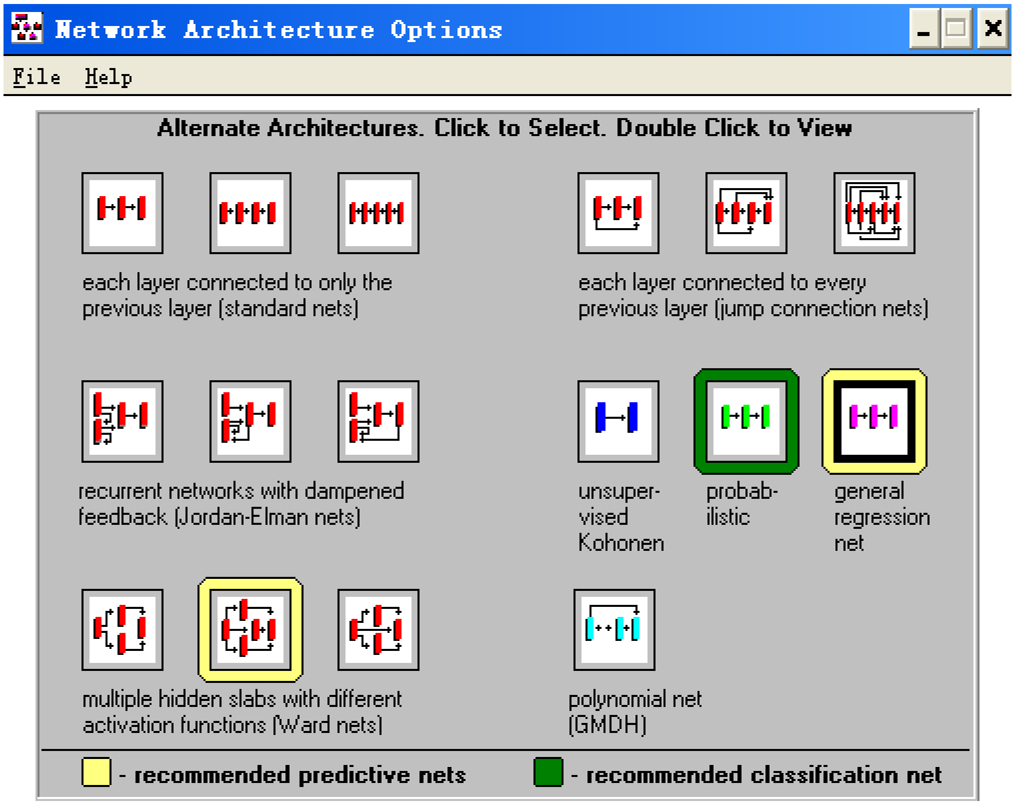
NeuroShell 2 provides us with five alternate architectures, as shown in Figure 1. They are Backpropagation (BP), Unsupervised (Kohonen), Probabilistic Neural Network (PNN), General Regression Neural Network (GRNN) and GMDH Network (Group Method of Data Handling or Polynomial Nets). Due to that Kohonen and PNN architecture require more than one output. Given that our training set has only one output. We therefore train our data set with other three architectures. In the process of selecting architectures, we make sure other parameters remain same. In BP architecture, we separately try three neural networks: standard nets with three layers, standard nets with four layers and jump connection nets with four layers. After analyzing accuracy rate of the test set and R squared. We decide to choose GRNN architecture (the process of result analysis is described in Section 4.1).

3.2 The selection of test sets amount

We also take the amount of test sets into consideration. In our experiment, we respectively extract 10 percent, 20 percent and 30 percent samples as our test sets in the GRNN architecture. As a result, test sets with 20 percent of all training sets are proved with the highest accuracy rate. Therefore, in our research, we select test sets with 20 percent to do our experiments (we provide concrete data analysis in Section 4.2).

3.3 The selection of neurons amount

After selecting neural network architecture and test sets amount, we conduct 12 experiments with different neurons in GRNN architecture. We begin the experiment with 50 neurons. Each time, we add neurons by 50. We find that the accuracy rate of 100 neurons and 200 neurons are the highest. Therefore, we also perform experiments respectively with 75 neurons, 125 neurons, 175 neurons and 225 neurons to explore a better result. The accuracy rate of these four experiments is no larger than the accuracy rate of 100 neurons and 200 neurons. At the same time, the R squared of 100 neurons is higher than 200 neurons and 225 neurons. We thus select 100 neurons as our candidate (the analysis process is illustrated in Section 4.3).



Alternate Architectures. Click to Select. Double Click to View

each layer connected to only the previous layer (standard nets)

each layer connected to every previous layer (jump connection nets)

recurent networks with dampened feedback (Jordan-Elman nets)

unsuper-vised Kohonen

probab-ilistic

general regression net

multiple hidden slabs with different activation functions (Ward nets)

polynomial net (GMDH)

🟨 - recommended predictive nets

🟩 - recommended classification net

Figure 1: Five neural network architectures in NeuroShell 2

4 Results and Analysis

4.1 The process of selecting architecture

As depicted in Section 3.1, the neural network architecture chosen in our research is GRNN architecture. This section will explain why we choose this architecture.

Table 1: Result of architectures selection

	BP			GRNN	GMDH
	Standard nets (three layers)	Standard nets (four layers)	Jump connection nets (four layers)		
R squared	0.5352	0.5559	0.5118	0.6180	0.3304
Accuracy rate	0.508772	0.508772	0.438596	0.614035	0.491228

Through analyzing the accuracy rate of neural network architectures, we notice that the accuracy rate of the GRNN is the highest among five architectures. In BP architecture, the accuracy rate of standard nets are higher than jump connection nets. Standard nets with three layers and four layers have the same accuracy rate. However, the R squared of standard nets with four layers is higher than three layers. We thus think that standard nets with four layers have better performance in classification and prediction.

4.2 Data analysis of selecting test sets

In Section 3.2, we select test sets with 20 percent of whole training sets as our samples. Through three experiments, we conclude the accuracy rate and R squared in Table 2.

Table 2: Result of test sets amount selection

GRNN architecture			
	Test set (10%)	Test set (20%)	Test set (30%)
R squared	0.4527	0.6180	0.5640
Accuracy rate	0.464286	0.614035	0.523256

As we can see in Table 2, both R squared and accuracy rate of test sets with 20% amount is higher than other two candidates. It is apparent for us to choose test sets with 20% amount as our test set.

4.3 Data analysis of selecting neurons

In Section 3.3, we choose 100 neurons in our research, the accuracy rate and R squared of 12 experiments are listed as follows:

Table 3: Accuracy rate and R squared of different neurons

GRNN architecture, 20% test set						
The number of neurons	**50**	**75**	**100**	**125**	**150**	**175**
R squared	0.66691	0.6670	0.7209	0.6802	0.5889	0.5982
Accuracy rate	0.54386	0.596491	0.649123	0.614035	0.561404	0.631579
The number of neurons	**200**	**225**	**250**	**287 (default by system)**	**300**	**350**
R squared	0.6013	0.5870	0.6180	0.6180	0.6180	0.6180
Accuracy rate	0.649123	0.649123	0.614035	0.614035	0.614035	0.614035

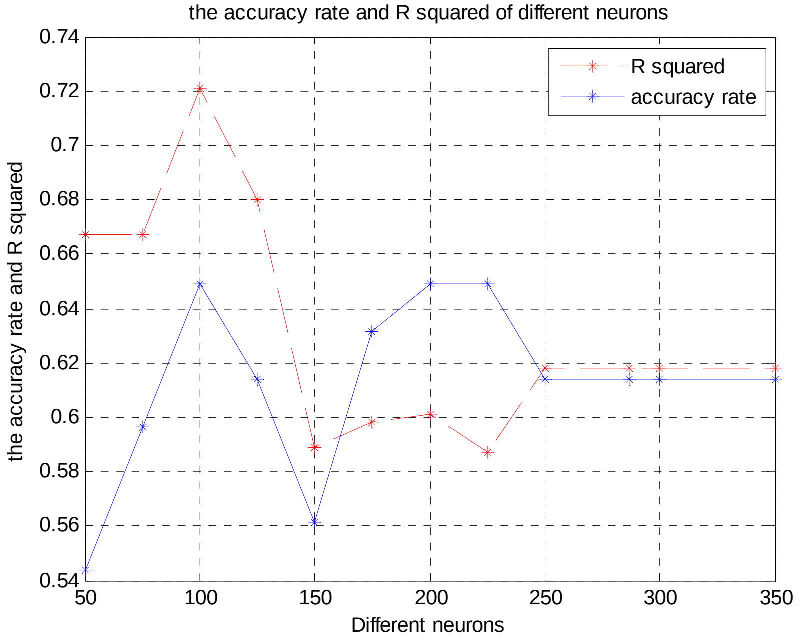


Figure 2: Accuracy rate and R squared of different neurons

We convert the R squared and accuracy rate in Table 3 into Figure 2. From Figure 2, we can clearly see that the accuracy rate of 100 neurons, 200 neurons and 225 neurons are the highest and the value of accuracy rate is same, but the R squared of each neurons are different. Compared with other two neurons, the R squared of 100 neurons is much higher. As a result, we decide to employ 100 neurons in our research.

5 Limitations

Although the accuracy rate of experimental data is improved by selecting neural network architecture, test set rate and neurons step by step. There are still some limitations in this paper, which need to be tackled in the future research. First, after retrieving raw data from UCI machine learning repository, the solution for missing data is directly remove the relevant instance from the samples, which is proved to be an inefficient data processing method. A new solution proposed for this limitation is that filling the missing data with mean value of the whole column instead of .removing from instances. Second, in the selection process of neurons, more times of experiments should be conducted around 100 and 200, which may result in a higher accuracy rate. Finally, considering the limited time, only three comparison experiments are performed. More comparison experiments should be conducted to get a higher accuracy rate.

6 Conclusion

To sum up, this paper achieves an accuracy rate increment through separately comparing neural network architecture, test set rate and neurons in the hidden layer. First, five experiments are performed to select a neural network architecture and GRNN architecture are selected in the paper when compare the accuracy rate of each neural network architecture. Then, based on the GRNN architecture, we conduct three experiments with different test set rate and the test set rate with 20 percent are chosen. Finally, to confirm neurons in the hidden layer of GRNN architecture, 12 experiments are conducted and the accuracy rate of 100 neurons and 200 neurons is the highest among all the neurons samples. Through comparing the value of R squared, we conclude that R squared value of 100 neurons is higher than 200 neurons. We therefore select 100 neurons in the hidden layer of GRNN architecture.

References

[1] Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40(4), 1086-1093.

[2] Ganesan, K. (2015). Classification Framework Based on C4. 5 Algorithm for Medicinal Data. International Journal of Computer Science and Information Security, 13(4), 63.

[3] Ming, L., Yi, W., Xiaogang, D., Qiucheng, S., & Jiawei, T. (2011). Improved BP algorithm and its application to intelligent diagnosis of coronary heart disease. Proceedings of 10th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), Vol. 4, 204-207.