Classification of Red Wine Quality Using Neural Networks

Hu Wanling (2819**), Zheng Xuan (2819****)**

Submission date: 2017-05-29

Southeast University-Monash University Joint Graduate School, Suzhou, China

Abstract

Nowadays, wine is enjoyed by more and more consumers all over the world. Red wine has a high nutritional value, and the main factors that determine its nutritional value are some quality indexes. The main quality indexes can be generally divided into sensory indexes and physiochemical indexes. According to the physical and chemical indicators of grapes and wines, quality inspector can evaluate the quality of all kinds of wines. However, identification of the liquor quality sometimes costs a lot of manpower and material resources. Our research aims to use neural networks to complete the classification of red wine qualities with several chemical features. According to the results of experiments displayed on the report, we could evaluate the quality of red wines.

Keywords: neural network, classification, GRNN, R square, accuracy rate

1. Introduction

In the previous research, the classification of wine quality is based on gas chromatography method [1] and chemical composition. This is a time-consuming and inefficient process. It has the positive significance for the production of the enterprise to classify the quality of red wine quickly and effectively. Therefore, it is necessary to study a fast and efficient classification method. Xu et al. [2] propose an improvement of backpropagation neural networks to implement the classification of red wine qualities. In this research, we try to apply the neural network in the classification of red wine quality. Through comparing the different result by choosing different condition and training the same original data set, we arrive at an optimal result. We choose R squared and Accuracy rate as a criterion for judging the outcome. In general, accuracy rate and R squared are positive relative. The process of comparison is divided into three steps, and the section four will give more detail about the process. After comparing the result, we get the highest data of R squared with 0.7833 and highest accuracy rate with 0.8854. As a result, we think the General Regression Neural Network (GRNN) architecture with 30% test set and 1599 neurons could the best classification method through 13 experiments. Finally, we describe the limitation about the process of classification with building a neural network model.

This paper can be divided into six sections. Section 2 state the database that we search from UC Irvine Machine Learning Repository. Section 3 describe the process of selecting neural network, test set and neurons for the final experiment. Section 4 displays the analysis of all the experiments. Section 5 propose two limitations of this paper and section 6 summarizes this research.

2. Data sets

In our research, we obtain the data sets about attributes of red wines from UCI machine learning repository wine quality data set [3]. The data set contains two databases, one is about white wines and another is about red wines. Both of them use 11 attributes to perform as attributes to judge the quality of red wines. We choose data from red wine, and range of the quality degrees in this database goes from 3 to 8, which is more concentrated than what in white wine data set. We think the criterion of classifying the quality of red wines is more reasonable and easily to studied on. The attributes that we take into consideration are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. The number of instances is 1599, we will choose an appropriate proportion of the instances as test sets in following experiments. All the experiments contain 11 inputs and an output that represent the quality of red wines. Each output has 6 states to show the different degrees of red wine qualities: 3,4,5,6,7 and 8. These number represent higher and higher quality levels developing gradually.

3. Training issues

3.1 The selection of architecture

We will emulate the classification using NeuroShell 2, which support our experiment with five optional architectures. They are displayed in Figure 1, named Backpropagation (BP), Unsupervised (Kohonen), Probabilistic Neural Network (PNN), General Regression Neural Network (GRNN) and GMDH Network (Group Method of Data Handling or Polynomial Nets). Because there is only one output in our training set, we abandon using the Kohonen and PNN architecture, which demand two or more output. We compare the BP, GRNN and GMDH with training our data set, and choose three kinds of BP architecture, which belongs to standard nets with three layers, jump connection nets with three layers and jump connection nets with four layers, in addition.

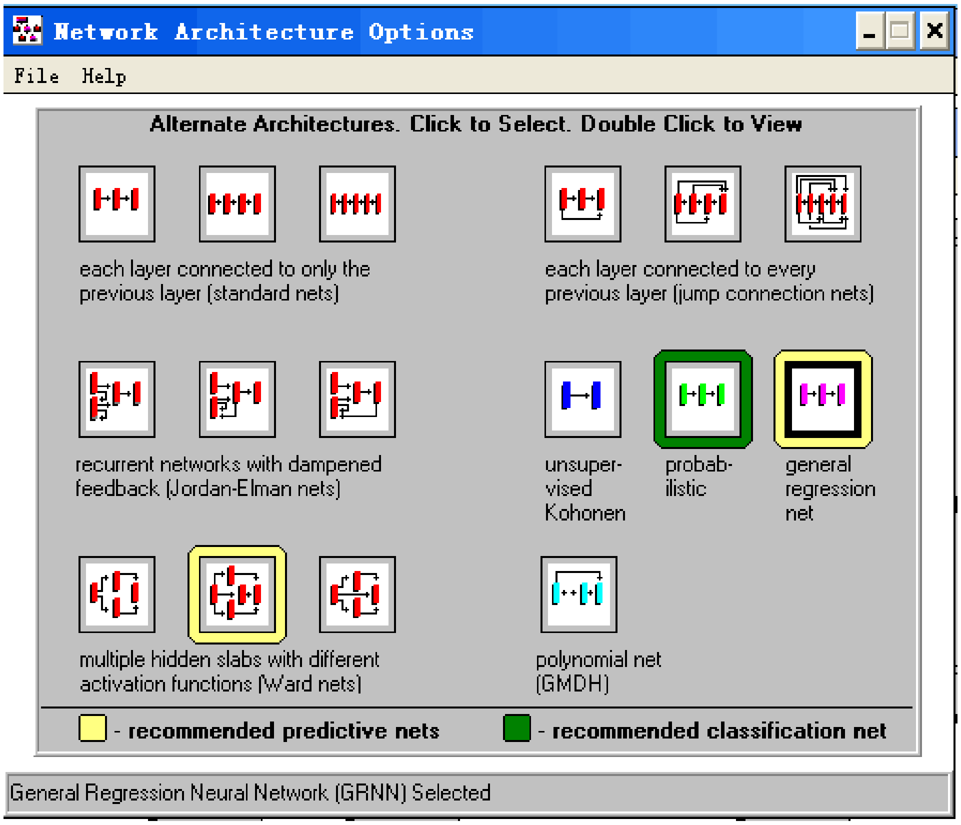
We carefully remain the other settings of the software are same to avoid introduce some unwanted outliers. We analyze and evaluate the R squared and accuracy rate of each test and make the decision of choosing GRNN as the final architecture we used in following experiments at last. Section 4.1 make a detailed analysis on the performances of architectures.

3.2 The selection of test sets amount

After determining the GRNN architecture, we then make a tradeoff among different amounts of test set. We successively extract 10 percent, 20 percent, 30 percent and 40 percent from our instances as the test sets, and compare the R squared and accuracy rate. Finally, we decide extract 30 percent from samples every time as the data set with 30 percent performs the highest accuracy. We will describe the process of our analysis roundly in Section 4.2.

3.3 The selection of neurons amount

As we already choose the architecture of GRNN and test set quantity of 30%, we implement the experiment for 8 times with diverse amounts of neurons. We set the initial number of neurons as 1599, increase and decrease the neurons with the unit as 200. The schematic diagram of the GRNN layers is shown in Figure 2. The least number of neurons is 999 and the biggest quantity is 2399. Just as the former two steps, we compare the R squared and accuracy rate of all these 8 experiments, then we draw a conclusion that the classification performs best when we choose 1599 neurons. From these 8 experiments, we discover the overall law of choosing suitable quantity of neurons and display the analysis refractive chart 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)

recurrent networks with dampened feedback (Jordan-Elman nets)

unsupervised Kohonen

probabilistic

general regression net

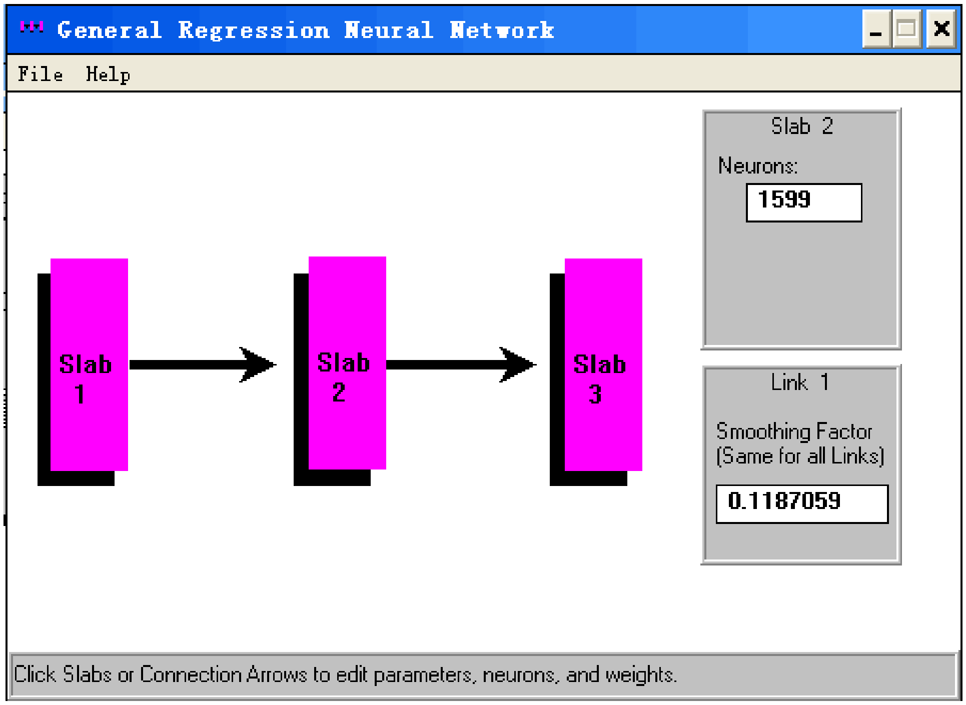
multiple hidden slabs with different activation functions (Ward nets)

polynomial net (GMDH)

🟨 - recommended predictive nets 🟩 - recommended classification net

General Regression Neural Network (GRNN) Selected

Figure 1: neural network architectures provided by NeuroShell 2



Slab 2

Neurons: 1599

Link 1

Smoothing Factor (Same for all Links)

0.1187059

Click Slabs or Connection Arrows to edit parameters, neurons, and weights.

Figure 2: Setting the neurons in GRNN architecture

4. Results and Analysis

4.1 The process of selecting architecture

In this experiment, we choose the GRNN architecture of neural network (see Figure 1). The R squared and accuracy rate for different condition of choosing are shown. This section will describe more details of selection process and data comparison.

Table 1: Result of architectures selection

	BP			GRNN	GMDH
	Standard nets (three layers)	Ward nets (two Hidden slabs)	Ward nets (three Hidden slabs)		
R squared	0.2508	0.2610	0.2765	0.5618	0.3970
Accuracy rate	0.603112	0.571875	0.581251	0.750001	0.62181

Through analyzing the results by training the original data set of red wine in different architectures and parameters, include BP, GRNN and GMDH architecture. It can be seen that GRNN architecture could get the highest data of R squared and accuracy rate (see table 1). The GMDH has the second highest R2 and accuracy rate. In the BP architecture, the accuracy rate of two hidden slabs and three hidden slabs match closely, and the three layers of standard nets has the highest R2 and accuracy rate among those three architectures. Therefore, we choose the GRNN architecture for our next work of classification, because it could have better performance in classifying the wine quality.

4.2 Data analysis of selecting test sets

In the comparison of different architecture, we select the default value as our samples. The default percent is 20%. Through training four different test set, include 10%, 20%, 30%, 40%, we can get the highest and lowest R2 and accuracy rate in table 2.

Table 2: Result of test sets in GRNN

GRNN architecture				
	Test set (10%)	Test set (20%)	Test set (30%)	Test set (40%)
R squared	0.4923	0.5618	0.7833	0.4813
Accuracy rate	0.69375	0.7500	0.8854	0.6687

In table 2, we can found that there are increasing and decreasing trend, and it reached a peak in 30% test set. In 30% test set, it has the highest R squared of 0.7833 and the highest accuracy rate of 0.8854. So, we choose the test set with 30 percent of whole training set for the next comparison.

4.3 Data analysis of selecting neurons

In the comparison of different architecture and test set, we all select the default value as our samples. The default number of neurons is 1599. Table 3 shows the results of choosing different number of neurons.

Table 3: Accuracy rate and R squared of different neurons

GRNN architecture, 30% test set				
The number of neurons	999	1199	1399	1599（default by system）
R squared	0.4971	0.5193	0.4628	0.7833
Accuracy rate	0.6896	0.6917	0.6708	0.8854
The number of neurons	1799	1999	2199	2399
R squared	0.5177	0.5209	0.5173	0.5099
Accuracy rate	0.6916	0.6875	0.6854	0.6775

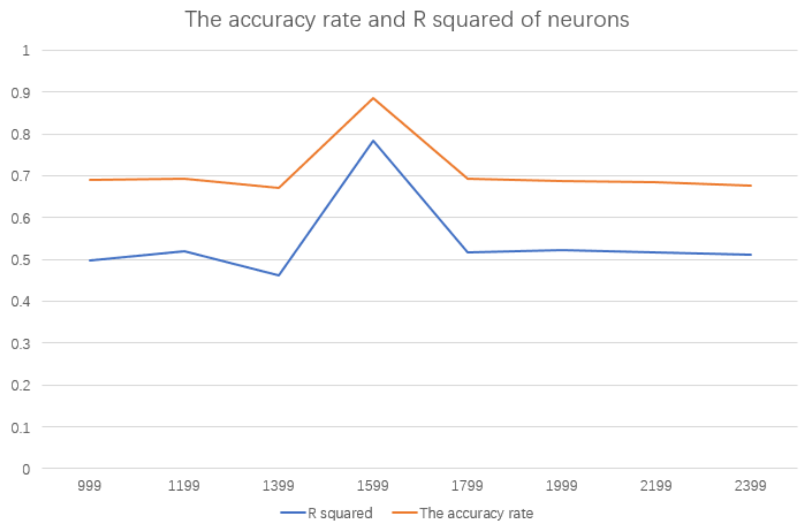


Figure 3: R squared and Accuracy rate of different neurons

We list the result of setting different number neurons in Table 3, and create a simple line graph to display the trend (see Figure 3). From Table 3, we can found that there are increasing and decreasing trend from the 999 neurons to 1799 neurons. It assumes an obvious growth tendency from 1399 neurons to 1599 neurons, and it becomes steadily smaller with the increasing of neurons from 1799. It is clearly that 1599 neurons have the highest result of R squared and accuracy rate. Therefore, we select the 1599 neurons in our research.

5. Limitations

We conduct our experiments through choosing applicable neural network architecture, test data and amounts of neurons orderly. But our research still has some limitations. First, inside the samples of our database should be more valuable attributes to measure the quality of red wines. And several of the attributes may be correlated. If time permits, we can search more data to summarize more features in order to evaluate synthetically. Second, in the section of selecting neurons, we should implement more experiments within the range from 1399 to 1799, as the accuracy rate drop abruptly in both side of 1599 and then has no obvious changes from 999 to 1399 as well as 1799 to 2399. Our experiments are designed to detect the rough changing trend of accuracy rate without considering details. The method for this limitation is to conduct more experiments with the neurons between 1399 and 1799, which may reveal more subtle changes about the classification performances

6. Conclusion

In conclusion, our research reaches an improvement of the accuracy rate by choosing the optimal neural network architecture, test set and the quantity of neurons. First, we implement five experiments to compare the performances of different architectures, with other settings remaining default values. We choose GRNN as the final one using in future researches. Second, we contrast three experiments under the conditions that the test sets are set as linear increasing percent of the samples based on GRNN architecture. At last, we complete 8 experiments for the decision of neurons that contained in the hidden layer of GRNN. We find that the classification gets the best effect with 1599 neurons. We use the R squared and accuracy as the indicators to determine the selection of each step.

According to this experiment, we come to a conclusion that Neural Network could perform a high accuracy in dealing with the classification of red wine quality. Among the architectures provided by NeuroShell 2, GRNN could offer the optimal classification with the shortest time and highest accuracy.

2023 Words

References

[1] Botelho, G., Mendesfaia, A., & Clímaco, M. C. (2010). Gcolfactometry and descriptive sensory analysis in the study of clonal red wines. *Ciência E Técnica Vitivinícola, 25*(1), 15-24.

[2] Xu, J., Ma, C., Zhang, Z., Luo, Y., & Liang, Z. (2015). Wine quality evaluation based on self-organizing feature map networks (SOM network) and BP neural network. *Journal of Yangtze University*.

[3] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*(4), 547-553.