Classification of Wine Using Neural Networks

Yuxin Zhang
yzha431@student.monash.edu
Southeast University – Monash University Joint Graduate School
Suzhou, China

Introduction

This paper aims to use Neural Networks to classify the cultivars of different types of wine with the chemical attributes of wine (i.e., wine recognition). According to the experiments which have been presented in this paper, we could determine which chemical attributes of wine are the feature chemical attributes of different wine cultivars.

Many related experiments have been conducted. Liu et al. [1] have studied the discrimination of rice wine age while using visible and near infrared spectroscopy combined with back propagation (BP) Neural Network. In their study, the recognition rate of the rice wine age is 96.67%. Sun et al. [2] have also studied the Chinese wine classification while using back propagation (BP) Neural Network. Their result of Chinese wine classification is also satisfactory. As we can see, Neural Network has become the ideal tool for solving this kind of problem. This is because Neural Network is capable of learning the features of the data. At the same time, Neural Network ensures the relationships discovered will generalize to new data. Thus, this paper will also use Neural Network to determine the cultivars of wine.

In this paper, a basic error backpropagation (BP) algorithm is employed in the training of the Multilayered Feedforward Neural Network (MFNN). BP neural network is one of the most popular Neural Network topologies [3][4]. The attributes will be processed by Neural Network and the network output expresses the resemblance that an object corresponds with a training pattern, i.e. the dissimilarity between the desired (i.e. 0 or 1) and calculated network output is employed to adjust the weight factors between the neurons. Along with every pass of a training pattern and adjustment of the weight factors, the difference between the desired and calculated network output, defined as the network output error, will gradually become less until it meets the desired value. One cycle through all the training patterns is defined as an epoch. In our study, the software called NeuroShell 2 will be used to process the data set while using the BP algorithm. Meanwhile, NeuroShell 2 offers different kinds of the architectures of Neural Network and parameters for users to choose. This paper will introduce several different experiments which have been conducted by different architectures and parameters.

Data Sets

This data set about the wine is obtained from the UCI Machine Learning Repository. Originally, these data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The data set is used to define the cultivars of wine with 13 chemical attributes. This data set contains 178 instances of wine attributes and their cultivars (e.g., region 1, region 2 and region 3) in Italy. 59 of these instances are belonged to region 1. 71 of these instances are belonged to region 2. The rest of the instances are belonged to region 3. Altogether, this data set has 13 numerical attributes and 3 classes (i.e. cultivars).

Since all attributes are continuous and numerical, we do not need to preprocess any attribute. However, the classes of wine are needed to be preprocessed. Because the Neural Network may perform better in doing classification when the Neural Network architectures have the same number of output neurons as the classes have. Then the region 1 will be represented as (0 0 1), the region 2 as (0 1 0) and the region 3 as (0 0 1). This technique is called 1-out-of-N encoding.

The detail of the attributes and classes are as follows:

ATTRIBUTES

1) Alcohol	numerical
2) Malic acid	numerical
3) Ash	numerical
4) Alcalinity of ash	numerical
5) Magnesium	numerical
6) Total phenols	numerical
7) Flavanoids	numerical
8) Nonflavanoid phenols	numerical
9) Proanthocyanins	numerical
10) Color intensity	numerical
11) Hue	numerical
12) OD280/OD315 of diluted wines	numerical
13) Proline	numerical

CLASSES

1)	REGION 1:	(0 0 1)
2)	REGION 2:	(0 1 0)
3)	REGION 3:	(1 0 0)

Training Issues

In Experiments 1 and 2, 13 chemical attributes have been chosen from the data set as inputs. Then, we use NeuroShell 2 to build a MFNN architecture which contains 13 inputs, 15 hidden layers and 1 output (i.e., region 1 is represented as 1, region 2 as 2 and region 3 as 3). In order to form the test set, 20 % of the data set has been randomly extracted. We set the learning rate and momentum factor fixed at c = 0.1 and α= 0.1. In the meantime, the initial weights are small random numbers around 0.3. The performance of this MFNN is measured on the test set every 200 epochs. The order of the inputs is random. In order to prevent memorization of the training data, the training will be terminated as long as the test set error had not improved within 1,000 epochs. For Experiments 3, 4 and 5, they have 4 attributes as input neurons. The number of Hidden Layer is 15, 5 and 25 respectively. Other parameter settings will remain the same as Experiments 1 and 2. In Experiments 6, 7 and 8, the input neurons are 6 attributes. The other parameter settings are corresponding to Experiment 3, 4 and 5 respectively. For Experiments 9, 10 and 11, the most of their parameter settings are corresponding to Experiments 3, 4 and 5 respectively. Experiments 9, 10 and 11 have 10 attributes as input neurons.

Results

All of the 11 experiments have been done to determine which chemical attributes are the feature chemical attributes of different wine from different cultivars. They are divided into 4 parts.

The first part is Experiments 1 and 2. The main difference between these two experiments is the output neurons. Experiment 1 has only one output neuron, while Experiment 2 has three. According to these two experiments, we could find out whether the number of output neurons will influence the feature chemical attributes or not. Tables 1 and 3 summarize the classification results.

Table 1 Classification accuracy of Experiment 1

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	71	0	100%
Actually REGION 3	0	0	48	100%
Column Accuracy	100%	100%	100%	

Table 2 Classification accuracy of Experiment 2

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	71	0	100%
Actually REGION 3	0	0	48	100%
Column Accuracy	100%	100%	100%	

Table 3 Top 10 chemical attributes of Experiment 1 and Experiment 2

	Experiment 1	Experiment 2
1	Flavanoids	Flavanoids
2	Proline	Proline
3	Color intensity	Color intensity
4	Alcohol	Alcohol
5	Ash	Diluted wines
6	Malic acid	Alcalinity of ash
7	Diluted wines	Hue
8	Proanthocyanins	Ash
9	Alcalinity of ash	Malic acid
10	Hue	Proanthocyanins

According to results of Experiments 1 and 2, the trained neural network model can be expected to accurately classify the cultivars of wine 100 % of the time. Then, looking at the contribution each attributes is making to the decision making process reveals that some attributes are performing extremely significantly to the output of this MFNN. Table 3 shows the top 10 most-contributed attributes of both experiments. We can see that the only difference between the top 10 attributes of each experiment is the order. Although the classification result is incredibly high, it, to some extent, shows the fact that the Top 10 attributes will not be influenced by the number of output neurons. Therefore, in the following experiments, the output neurons will remain as 3.

The second part is Experiments 3, 4 and 5. All of these three experiments are conducted with the top 4 attributes, Falvanaoids, Proline, Color intensity and Alcohol, as the input neurons of the MFNN. The main difference of these three experiments is the number of Hidden Layers. Firstly, we choose to have 15 Hidden Layers. Table 4 shows the result of classification.

Table 4 Classification accuracy of Experiment 3

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	70	1	98.59%
Actually REGION 3	0	1	47	97.92%
Column Accuracy	100%	98.59%	97.92%	

According to the Table 4 the result of classification is still relatively high, comparing to the Experiments 1 and 2, these 4 attributes are still not capable of representing the feature attributes of the wine from different cultivars. Therefore, we start doing the Experiments 4 and 5 by changing the number of Hidden Layers to see whether the number of Hidden Layer is the key factor that influence the result. Tables 5 and 6 show the results of Experiments 4 and 5 respectively.

Table 5 Classification accuracy of Experiment 4

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	58	1	0	98.31%
Actually REGION 2	1	69	1	97.18%
Actually REGION 3	0	1	47	97.92%
Column Accuracy	98.31%	97.18%	97.92%	

Table 6 Classification accuracy of Experiment 5

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	70	1	98.59%
Actually REGION 3	0	1	47	97.92%
Column Accuracy	100%	98.59%	97.92%	

According to the Tables 5 and 6, the accuracy of classification of Experiment 4 is relatively lower than Experiment 3. On the other hand, Experiment 5 shows the same result as Experiment 3. Therefore, these four attributes are not capable of representing the feature attributes of wine from different cultivars.

The third part is Experiments 6, 7 and 8. In Experiment 6, we add more attributes as inputs.

However, according to Table 3, the order of the top 10 attributes is no longer the same. Therefore, in the top 8 attributes, we choose the 6 attributes which in the both results of Experiments 1 and 2 as the input neurons of this MFNN. The 6 attributes are Falvanaoids, Proline, Color intensity, Alcohol, Diluted wines and Ash. Table 7 shows the result of Experiment 6.

Table 7 Classification accuracy of Experiment 6

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	70	1	98.59%
Actually REGION 3	0	0	48	100%
Column Accuracy	100%	100%	97.96%	

As we can see from Table 7, the result is becoming better than Experiment 3. However, due to the perfect performance of Experiments 1 and 2, we still cannot say these 6 attributes is capable of representing the wine from different cultivars. Experiments 4 and 5 have also indicated that the Hidden Layers may influence the result of classification. Therefore, we continue to conduct Experiments 7 and 8. Tables 8 and 9 show the results of Experiments 7 and 8 respectively.

Table 8 Classification accuracy of Experiment 7

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	58	1	0	98.31%
Actually REGION 2	0	70	1	98.59%
Actually REGION 3	0	0	48	100%
Column Accuracy	100%	98.59%	97.96%	

Table 9 Classification accuracy of Experiment 8

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	70	1	98.59%
Actually REGION 3	0	0	48	100%
Column Accuracy	100%	100%	97.96%	

The results of Experiments 7 and 8 reveal the fact that the changes of Hidden Layers are not able to make the accuracy better when the inputs are these 6 attributes. Thus, these 6 attributes are still not what we are looking for. They cannot be represented as the feature attributes.

The fourth part is Experiments 9, 10 and 11. Looking at the results of last experiments, 6 attributes are still not enough. Then we decide to add all of the Top 10 attributes as the input neurons. The MFNN architecture of Experiment 9 is 10-15-3. Table 10 shows the result.

Table 10 Classification accuracy of Experiment 9

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	71	0	100%
Actually REGION 3	0	0	48	100%
Column Accuracy	100%	100%	100%	

Eventually, we have a satisfactory result as Experiments 1 and 2. Then Experiments 10 and 11 need to be done to eliminate the interferential factor. In the Experiments 10 and 11, the number of Hidden Layers has been changed to 5 and 25 to see whether the result changes or not. Tables 11 and 12 show the results that the number of Hidden Layers will not influence the accuracy of classification with these 10 attributes as input neurons.

Table 11 Classification accuracy of Experiment 10

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	71	0	100%
Actually REGION 3	0	0	48	100%
Column Accuracy	100%	100%	100%	

Table 12 Classification accuracy of Experiment 11

	Classified REGION 1	Classified REGION 2	Classified REGION 3	Row Accuracy
Actually REGION 1	59	0	0	100%
Actually REGION 2	0	71	0	100%
Actually REGION 3	0	0	48	100%
Column Accuracy	100%	100%	100%	

Therefore, we can finally draw the conclusion that these 10 attributes, Falvanaoids and Proline, Color intensity and Alcohol, Diluted wines and Ash, Alcalinity of ash and Hue, Proanthocyanins and Malic acid, are capable of representing the feature attributes of wine from different cultivars.

Limitations

Due to size of the data set, the result is still convincible. Although 11 experiments have been done, there are lots of parameters have not been examined. Meanwhile, there are many other attributes could be significant to the results. Moreover, instead of Neural Network and BP algorithm, many other techniques have not been tried to do the classification of this subject (e.g., Nearest-Neighbor Classifiers and Decision Trees). In the future studies, more experiments are needed to be done to ensure a better performance of the classification of wine from different cultivars.

Conclusion

This paper presents the classification of wine from different cultivars using Neural Networks. NeuroShell 2 has been used to build the model of MFNN architecture. According to the results of 11 experiments, we finally have the 10 feature attributes to represent the wine from different cultivars. Although, the size of the data set is too small to have a relatively convincible result. Due to the insufficient instance, the accuracy of the classification may be too high to believe. This paper still shows the feasibility to do the classification using Neural Networks. At the same time, the Neural Network model shows its capability as the ideal tool of classification.

References

[1] F. Liu, F. Cao, L. Wang, and Y. He, "Discrimination of Rice Wine Age Using Visible and Near Infrared Spectroscopy Combined with BP Neural Network," *IEEE 2008 Congress on Image and Signal Processing*, pp. 267-271, 2008.

[2] X. Sun, X. Tang, and Y. Lei, "Continuous attribute discretization and application in Chinese wine classification using BP neural network," *IEEE 2008 3rd International Conference on Intelligent System and Knowledge Engineering*, pp. 896-900, 2008.

[3] V. Kecman, *Learning and Soft Computing, Support Vector Machines, Neural Networks and Fuzzy Logic Models*, The MIT Press, Cambridge, MA, 2001.

[4] L. P. Wang, X. J. Fu, *Data Mining with Computational Intelligence*, Springer, Berlin, 2005.