PAPER • **OPEN ACCESS**

The Mathematical Analysis and Classification Research of an Iris Data Set Using Binary Tree and Grey Relation Grade

To cite this article: Chiang Ling Feng 2021 *J. Phys.: Conf. Ser.* **2068** 012004

The Mathematical Analysis and Classification Research of an Iris Data Set Using Binary Tree and Grey Relation Grade

Chiang Ling Feng*

Department of Electric Engineering, ChienKuo Technology University, Taiwan

*Corresponding author email: acclaim0629v@gmail.com

Abstract. The data from an Iris flower database is studied. The Iris database is the most commonly used database for machine learning algorithms. The Iris database was developed by Ronald Aylmer Fisher in 1936. The Iris database has 150 records in three categories: Iris Sentosa, Iris Versicolor and Iris Virginic. The database has four attributes: sepal length, sepal width, petal length and petal width. For the machine learning algorithm, 150 Iris flower databases are used. Of the 150 Iris in the Iris database, 80% are used as the training set and the remaining 20% Iris as the test set. In machine learning, to perform classification and discrimination is a complicated and difficult thing. In this study, a grey relation grade is used to extract the main features of the Iris flower and a Binary Tree [1] is used to classify the Irises. The results show that for the same specific attributes, grey relation grade extracts the main attributes and can be used in combination with a binary for classification.
Keywords: Grey theory, grey relation grade, binary tree.

1. Introduction

A back propagation neural network BPN uses forward propagation and backward propagation. Forward propagation feeds training data into the network to calculate the error between the result and the corresponding target. Backward propagation adjusts the network weight according to the error value that is calculated using forward propagation, in order to correct the network to give a result that is within the smallest error range after many training cycles. In one study [2], the parameters are first initialized and the weights and biases are randomly established. The hidden layer node is 6 and the maximum number of iterations is 10,000. This study determined that there are four data classification errors. After further adjustment of the model, the training accuracy was improved. The disadvantage of this method is that there is overfitting and the calculation time is very long. Another study [3] achieves an accuracy of only 96.67% using a hidden layer node of 3, 5000 repeated operations and a learning rate of 0.01. Using a neural network, the classification of the Iris database requires a very low learning rate and a large number of repeated operations. This is closely related to the number of hidden nodes and the training data set and the test data set. The control factors are also too complicated. T One study [4] that uses the data from the Iris flower database shows that the sepal width and sepal length of these 150 Iris are not correlated, but the petal width and petal length are highly correlated. This study uses grey theory to verify the results of this study [4].

The nearest neighbor method uses the vector space model to classify. The concept is that the cases of the same category have a high degree of similarity to each other. The possible classification of cases of unknown categories can be evaluated by calculating the similarity with cases of known categories. The

disadvantage of the k-nearest neighbor algorithm is that it is very sensitive to the local structure of the data.

Grey relation grade is a quantitative analysis method that is used to compare the relationship between data series. It reflects the degree of relationship between the essential curves of the data series. If the relationship between a data series and a reference data series is larger, the development trend for the data series has a closer relationship with the reference data sequence.

This study uses 150 data points from the Iris flower database and the fuzzy characteristics and classification of uncertain attributes is used. The paper is organized as follows. In the second section, we introduce the grey relation grade method. In the third section, we analyse the characteristics of Iris flowers among these 150 Iris data. We found the import influence factors among these characteristics using grey relation grade method. In fourth section, we also use the grey relation grade method combined with the classification method of binary tree to perform the flower classification. Conclusions are drawn in section five.

2. Grey Relational Grade

Grey relational analysis [5, 6] is a quantitative measurement method for a dynamic process in a system. It measures the degree of convergence between the entire data set using the development trend for multiple attributes. If the original data sequence has *m* attributes and each attribute original data sequence has *n* element values, then:

$$\begin{aligned} X_1^{(0)} &= \left\{x_1^{(0)}(1), x_1^{(0)}(2), \dots\dots, x_1^{(0)}(n)\right\} \\ X_2^{(0)} &= \left\{x_2^{(0)}(1), x_2^{(0)}(2), \dots\dots, x_2^{(0)}(n)\right\} \\ &\vdots \\ X_m^{(0)} &= \left\{x_m^{(0)}(1), x_m^{(0)}(2), \dots\dots, x_m^{(0)}(n)\right\} \end{aligned} \tag{1}$$

Grey correlation uses the proximity of the curve for each data set to determine whether the connection between each is close and which attributes are main influencing factors or secondary influencing factors. There is usually insufficient information so it is difficult to determine the relationship between attribute variables using only the data. The largest correlation between each factor is determined by calculating the relationship between each factor. To determine the degree of correlation for the data set, the correlation coefficient is calculated. The data sequence unit for each attribute in the data set may be different so the measurement unit must be determined before calculating relevance. Let

$$X_i^{(1)} = \left\{\frac{x_i^{(0)}(k) - \min\limits_k x_i^{(0)}(k)}{\max\limits_k x_i^{(0)}(k) - \min\limits_k x_i^{(0)}(k)}, k = 1,2, \dots\dots, m\right\} \tag{2}$$

$X_1^{(1)}$ tis the reference sequence and other data sequences are comparison data sequences. Define

$$\Delta_i(k) = \left|x_1^{(1)}(k) - x_i^{(1)}(k)\right|, i = 1,2, \dots\dots, m \tag{3}$$

and

$$V_{max} = \max_i \max_k \Delta_i(k) \tag{4}$$

$$V_{min} = \min \min_k \Delta_i(k) \tag{5}$$

Then, the correlation coefficient can be defined as

$$\gamma_{1,i} = \frac{V_{min} + \zeta V_{max}}{\Delta_i(k) + \zeta V_{max}} \tag{6}$$

Where ζ is the identification coefficient with a value between 0 and 1. The value of coefficient ζ is usually 0.5 so the correlation coefficient between attribute $i = 2,3, \dots\dots, m$ and attribute 1 can be

calculated. If $X_j^{(1)}$ is used as the attribute reference data sequence, the correlation coefficient between it and other attributes can also be calculated.

3. The Classification of Iris Data by Binary Tree

Using the classification algorithm of binary tree, we must first find out the targets that are easier to classify. After finding out, we then analyze the important influencing attributes of these goals. After finding these important attributes, you can find the attributes and critical values that can easily divide the remaining targets into different binary trees. This step continues until all important attributes have been traversed. After that, look for all possible attributes, and give these attributes a weighting factor to add as the critical value for its segmentation target. This algorithm is give in Figure 1.

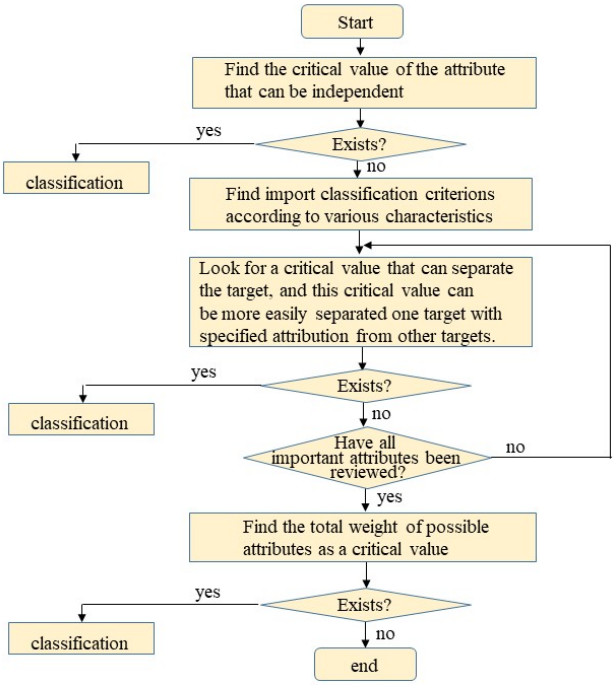

Figure 1. The generalized algorithm for binary tree.

The Iris Flower dataset [7] contains three species: Iris Sentosa, Iris Versicolor and Iris Virginica. Each type of Iris flower has four distinguishing features: sepal width, sepal length, petal width and petal length. Each species of iris flower features a specific range of sizes but not all parameters are used as features to create a decision tree for the available data using these groupings. This is because there are two reasons for making the decision that not all are used as features. First, from the petal features, both the width and length indicate the range that can be used to differentiate between types of Iris Sentosa, Iris versicolor and Iris Virginica. Second, the sepal feature shows a number that does not provide a significant differentiating number, this is because between species have almost the same range, namely the sepal length for Iris Sentosa species is in the size range between 4.3 - 5.8, versicolor 4.9 - 7.0, and Iris Virginica species between 4.9. - 7.9. Based on the size of the sepal width, the type of Iris Sentosa has a size range between 2.3 - 4.4, versicolor 2.0 and 3.4 and the type of Iris Virginica between 2.2 - 3.8. This shows that with this range it could be a certain range between the two types of Iris flowers. For example, the size of the sepal length is 5.0, it could be between the three types of Iris flowers, so it requires other parameters to determine the difference. See Table 1.

Table 1. Data range of various characteristics of each type of Iris.

	sepal length	sepal width	petal length	petal width
Iris Setosa	4.3~5.8	2.3~4.4	1.0~1.9	0.1~0.6
Iris Versicolor	4.9~7.0	2.0~3.4	3.0~5.1	1.0~1.8
Iris Virginica	4.9~7.9	2.2~3.8	4.5~6.9	1.4~2.5

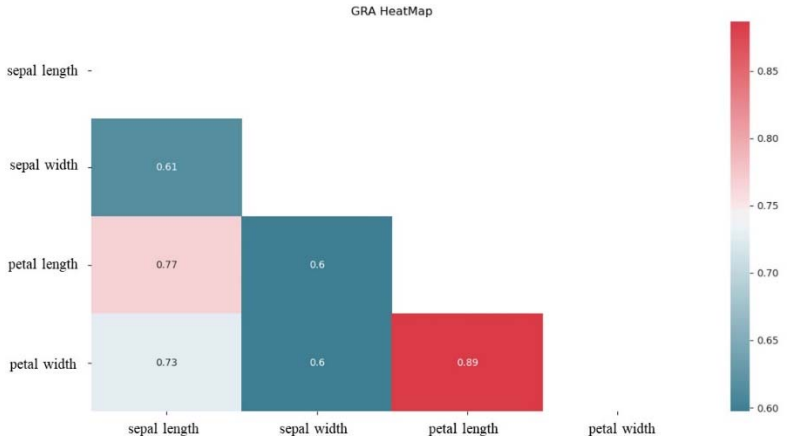

Figure 2. Iris database correlation coefficient.

From Figure 2, we can see that the characteristics of iris flowers can be determined using the petal length and the petal width. If the petal width is not greater than 0.6, Iris Sentosa can be distinguished from the other two Iris species. However, the sepal features have values in a similar range so cannot be used to differentiate between the three species. The sepal length for Iris Sentosa is 4.3 - 5.8, for Iris Versicolor is 4.9 - 7.0 and for Iris Virginica is 4.9. - 7.9.

Iris Sentosa has a sepal width of 2.3 - 4.4, Iris Versicolor has as sepal width of 2.0 - 3.4 and Iris Virginica has as sepal width of 2.2 - 3.8. These values can be used to differentiate between two types of iris flowers. An iris with a sepal length is 5.0 it could belong to any of the three species, so other parameters are required for differentiation. Figure 2 shows that petal width is one of the most informative attributes, so this is the first classification criterion because there are significant differences between each iris species in terms of this measure. From Table 1, if the petal width is less than 0.6, it is classified as Iris Sentosa type. If the petal width is greater than 1.8, it belongs is classified as Iris Virginica. For a petal width of 1.8, Table 10 shows that the species can be differentiated using the sum of the sepal length and the petal length. The decision binary tree to determine the species of iris is shown on the right side of Figure 3.

If the petal width is not greater than 1.8, there are 49 types of Iris Versicolor and 5 types of Iris Virginica. Using the attributes of these data, Iris Versicolor can be clearly distinguished if the petal length is less than 4.9. For other Iris species with a petal width of not less than 4.9, the related characteristics is classified as follows. If the petal length is not less than 5.6, Iris Virginica can be distinguished. Irises with a petal length of less than 5.6 are classified as Iris Virginica if the petal width is not greater than 4.5. If the petal width is not equal to 1.5, the species is classified as Iris Versicolor; otherwise, the species is determined using the grey theory correlation test. The final binary tree is shown in Figure 3.

4. Numeric Result

In the Iris data sheet, Iris is divided into three species of Iris Sentosa, Iris versicolor and Iris Virginica. Each species of Iris has four parameters as distinguishing characteristics, which are based on sepal length, sepal width, petal length and petal width. After calculating the petal width petal width, it is found that there are 10 data for the second category with petal width = 1.5 and 2 data for the third category with petal width = 1.5. Therefore, when the petal width of Iris is 1.5, it can be classified into Iris Versicolour and Iris Virginica. The classification of petal width of petal width = 1.5 is ambiguous. Therefore, we study the classification of Iris flowers by grey relation grade using the attribute petal width of petal width = 1.5 in this section.

According to the grey relation grade method, the data in Table 2 is formed under the condition of petal width = 1.5. It can be seen from Figure 4 that the degree of correlation between the petal length and the category attributes of Iris is 0.78.

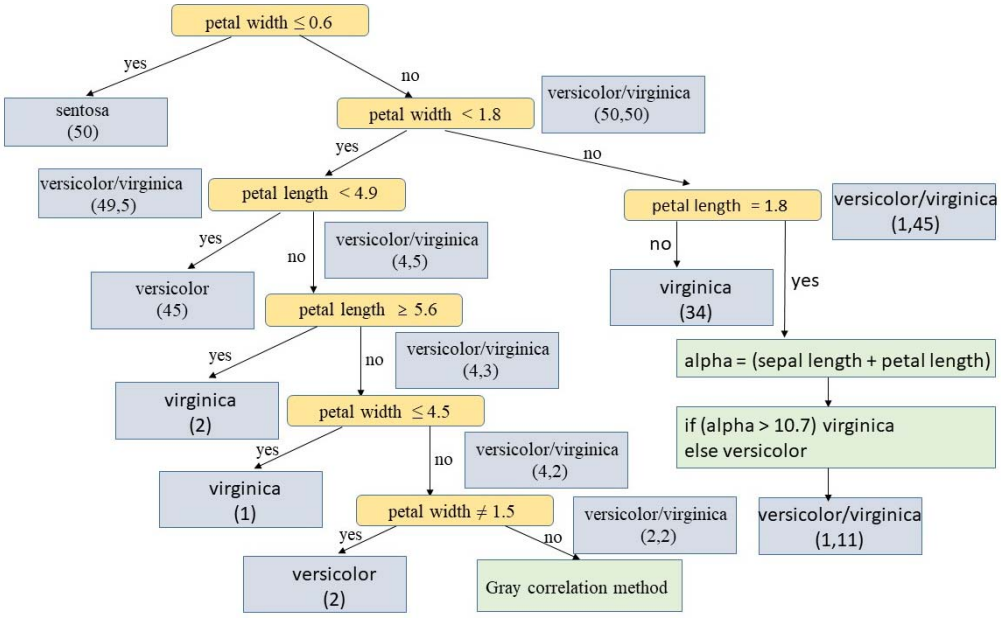

Figure 3. The complete Binary tree classification map of Iris.

Table 2. Attribute relationship table of Iris petal width=1.5.

item	sepal length	sepal width	petal length	petal width	species
1	6.4	3.2	4.5	1.5	Iris Versicolor
2	6.9	3.1	4.9	1.5	Iris Versicolor
3	6.5	2.8	4.6	1.5	Iris Versicolor
4	5.9	3	4.2	1.5	Iris Versicolor
5	5.6	3	4.5	1.5	Iris Versicolor
6	6.2	2.2	4.5	1.5	Iris Versicolor
7	6.3	2.5	4.9	1.5	Iris Versicolor
8	6	2.9	4.5	1.5	Iris Versicolor
9	5.4	3	4.5	1.5	Iris Versicolor
10	6.7	3.1	4.7	1.5	Iris Versicolor
11	6	2.2	5	1.5	Iris Virginica
12	6.3	2.8	5.1	1.5	Iris Virginica

Based on Table 2, set the petal length = 5 Iris flower category attribute to Iris versicolor. It can be seen that the correlation degree between the petal length and the category attributes of Iris is 0.75. Based on Table 2 the petal length = 4.9 Iris flower category attribute is classified as Iris Verginica. It can be seen that the relation grade between the petal length and the Iris category attribute is 0.71.

According to the above research, based on the condition of petal width = 1.5 and the Iris with petal length = 5 is classified into Iris Virginica, the correlation degree is 0.78. Based on the condition of petal width = 1.5 and the Iris with petal length = 5 is classified into Iris Versicolor, the correlation degree is 0.75. Based on the condition of petal width = 1.5 and the Iris with petal length = 4.9 is classified into Iris Virginica, the relation grade is 0.71. Therefore, the most suitable classification is that the Irises with petal length = 5 and petal length = 5.1 belong to Iris Virginica, and the Irises with petal length = 4.9 belong to Iris Versicolor (grey relation grade changed from 0.71 to 0.78).

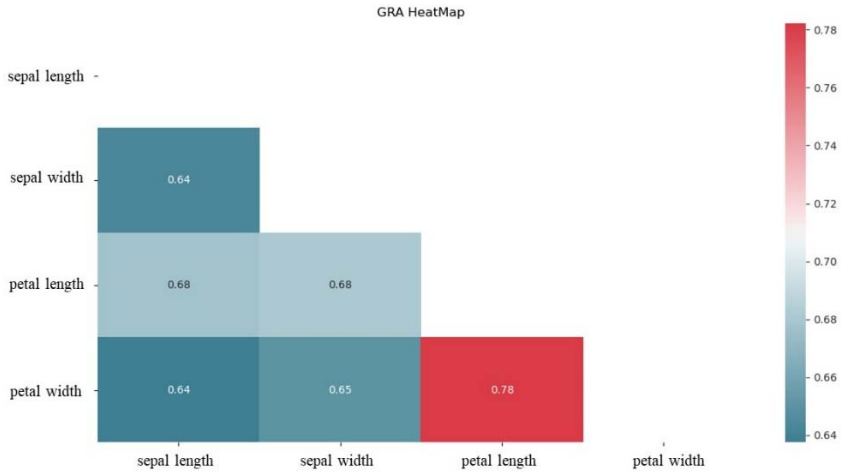

Figure 4. The relation grade table of the Iris database in Table 2.

5. Conclusion

From the discussion of the above examples, it is difficult to assign a main attribute for the Iris data. In the Iris Setosa variety, it can be clearly separated from the other two varieties by the petal width. As for Iris Versicolor and Iris Virginica, the attribute values of the two varieties overlap each other. For iris varieties with petal length = 1.5, we found that the two attributes of petal length and petal width can be used to separate iris varieties through the grey correlation method. For iris varieties with petal length = 1.8, the critical value can be set to the sum between sepal length and petal length through the algorithm shown in Figure 1 to easily separated iris varieties.

References

[1] Adrijan Božinovski and Nevena Ackovska, "The Binary Tree Roll Operation: Definition, Explanation and Algorithm", International Journal of Computer Applications (0975 – 8887) Vol. 46, No.8, pp. 40-47, May 2012.
[2] https://blog.csdn.net/weixin_33724659/article/details/89700732?utm_medium=distribute pc_relevant_download.none-task-blog-BlogCommendFromBaidu-22.nonecase&depth_1-utm_source=distribute.pc_relevant_download.none-task-blog-BlogCommendFromBaidu-22.nonecas.
[3] https://www.jianshu.com/p/52b86c774b0b.
[4] https://github.com/codingXiang/BPN.
[5] Ke H.F, and Chen B.L., "Novel gross error detection approaches of small samples based on *GM*(1,1) model, Vol. 25(1), pp. 44-53, 2013.
[6] Liu S.F., and Lin Y., Grey Systems: Theory and Applications, Berlin: Springer-Verlag, 2011.
[7] https://www.kaggle.com/saurabh00007/Iriscsv.