Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/37/20230467

Analysis of Naive Bayesian and Back Propagation algorithms in iris classification

Chengyang Yu

Department of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, 116024, China

20211072012@mail.dlut.edu.cn

Abstract. An efficient taxonomy of irises can provide botanists with valuable tools. Machine learning algorithms can effectively improve the performance of iris classification models because they can automatically analyze and summarize data. To this end, this paper introduces Naive Bayesian (NB) and Back Propagation (BP) to build classification models. When creating the NB model, the petal and sepal data from the iris dataset are used sequentially as classification criteria to classify the data. When constructing the BP model, the author sets different iterations and outputs the loss function and accuracy of the BP model under different iterations. The study finds that the NB model has higher classification accuracy when using petal length and petal width as classification criteria, which is 17% higher than the classification accuracy using sepal length and sepal width. Therefore, the NB model is more suitable for classifying independent data. By studying the use of the BP algorithm to classify iris flowers, the automatic classification of iris flowers can be realized and the accuracy of classification can be improved. Compared with the traditional NB algorithm, the BP algorithm can better mine the hidden patterns and information in the iris data and make effective classifications. This study provides new insights and discoveries for the taxonomic study of Iris plants.

Keywords: iris classification, naive bayesian, back propagation.

1. Introduction

Botanical identification and classification refer to the comparison and identification of an unknown plant species with known plant species by observing and analyzing their morphological, biological, and genetic characteristics, to determine their classification levels such as family, genus, and species, and provide accurate species names. Iris recognition and classification is an important research topic in botany recognition and classification.

Artificial Neural Networks (ANN) is a cutting-edge subject that has been extensively investigated in recent years. It is made up of a sizable number of linked artificial neurons. Based on biological functionality, it models how the biological nervous system functions [1]. As ANN has continued to advance, it has found widespread use, particularly in the field of classification. At present, there are nearly 40 neural network models. Among them, Back Propagation (BP) and Naïve Bayesian (NB) are commonly used models. The BP can flexibly handle linear and nonlinear patterns and has good recognition and classification performance [2]. The NB classification method is widely used, especially in the field of data knowledge mining [3]. By comparing the effectiveness of BP and NB in classifying

DOI: 10.54254/2755-2721/37/20230467

iris datasets, researchers can gain a deeper understanding of the functions and differences between these three models. At the same time, it also provides a new method for the classification of botany and helps to deeply explore the laws of plant classification has important scientific research and application value.

Before machine learning was well developed, botanists mainly used plant classification key tables to achieve classification and recognition of plant species [4]. This method can indeed accurately classify plants, but it is time-consuming and laborious. With the development of ANN, there are many iris classification methods based on deep learning algorithms, aimed to enhance accuracy and efficiency. To categorize the iris dataset in 2017, Amit Pandey, Achin Jain, et al. employed Min-Max normalization for various values of K [5]. The average precision for Min-max normalization was 88.0925% and Z-score normalization at 78.5675% [5]. In 2019, Yuanyuan et al. used the Random Forests model to classify the iris dataset. The accuracy of Boosting Tree was 60.9% while the accuracy of Random Forest was nearly 100% [6]. In 2020, Debaraj Rana attempted Principal Component Analysis (PCA) to classify the iris dataset. The accuracy of PCA was 86% [7]. In the same year, Zahraa et al. managed to use a Support Vector Machine (SVM) classifier with a PCA algorithm for feature reduction to classify the iris dataset [8]. All these features above show that ANN has extremely important significance in the classification of iris flowers as well as the development of Botanical Classification.

The major goal of this project is to build an NB and BP-based classification model for iris. Specifically, first, the Iris dataset is our main research objective. Second, the author divided the data into two parts: training data and testing data, with a ratio of 9:1. Then, the author used NB and BP to establish classification models for the dataset, respectively. The Bayesian algorithm has obvious advantages in classification. The NB classification method is widely used, especially in the field of data knowledge mining. Good stability and high efficiency are the core advantages of NB algorithms [9]. BP is a nonlinear mapping method with good recognition and classification performance [10]. Therefore, NB and BP can successfully classify iris flowers. When using NB to classify iris datasets based on sepal length and sepal width, the accuracy rate is 79.8%, and the accuracy rate is 96.8% based on petal length and petal width. When using BP classification, the accuracy is 100%. Through this study, the automatic classification of iris flowers can be achieved, and the accuracy of classification can be improved. At the same time, neural network algorithms can mine hidden patterns and information in iris data, provide new insights and discoveries for Taxonomy research of iris and other plants, and promote scientific research progress.

2. Methodology

2.1. Dataset description and preprocessing

The iris dataset, which Fisher gathered and compiled in 1936, is frequently used. The dataset is made up of 150 records that are split into 3 groups, each of which has 50 records and 4 attributes. Sepal length, sepal width, petal length, and petal width are the four identities that can be used to determine the kind of iris (Setosa, Versicolor, Virginia) [11]. The dataset consists of 5 columns: The first shows the length of the sepal, the second shows the width of the sepal, the third shows the length of the petal, the fourth shows the width of the petal, and the last shows the type of flower. The goal is to separate the data into four groups based on the dimensions of Iris flowers. The Iris flowers come in three different varieties, each of which is identified by a number from 0 to 2: Setosa (0), Versicolor (1), and Virginia (2).

2.2. Proposed approach

The focus of this research is to compare the classification performance of NB and BP on the Iris dataset. The fundamental idea of NB is to find the probability of each category under the condition that several items to be classified are known, and the category corresponding to the item with the highest probability. When using NB, this experiment sequentially uses the first two eigenvalues and the last two eigenvalues as training data. Firstly, the author imports the data. Secondly, calculate the Prior probability. The third step is to calculate the Conditional probability. Then, for the given test data x, calculate the Posterior probability of x. After calculating the maximum posterior probability and classifying instance x based

DOI: 10.54254/2755-2721/37/20230467

on the value of the maximum posterior guess, the process is complete. Figure 1 depicts the basic steps of the NB algorithm.

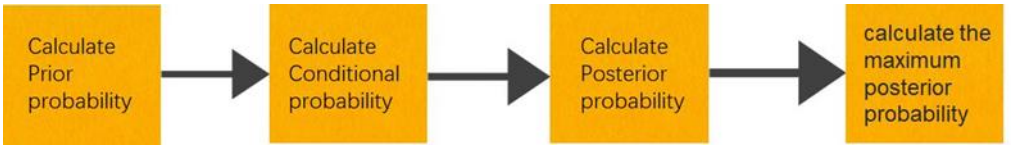

Figure 1. The pipeline of the NB model.

The main characteristics of the BP model: Firstly, the signal propagates forward, with input signals undergoing sequential processing through hidden layers before ultimately reaching the output layer. The second is error backpropagation. If the output of the output layer differs from what was expected, calculate the output error and send the error signal back via the original connection pathways. Once the intended goals are attained, each layer's neurons' weights and thresholds are adjusted. Figure 2 depicts the BP neural network method in action.

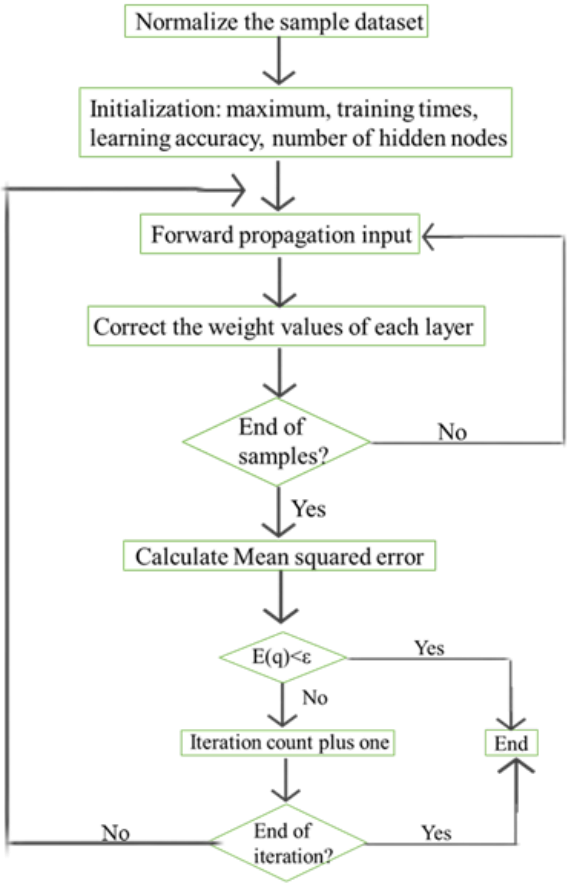

Figure 2. The main process of the BP model.

2.2.1. NB model. A comparable reduction based on the Bayesian method is the NB technique, which assumes attributes are Conditionally independent of one another when the target value is supplied. It has four steps: Prior Probability. In the iris dataset, the prior probability is the percentage of each species. The formula:

$$P(Y = C_k) = \frac{\sum_{i=1}^{N}(y_i = C_k)}{N}, k = 1,2, \ldots, K \quad (1)$$

DOI: 10.54254/2755-2721/37/20230467

In the formula (2), N is the total number of samples, and y_i is the i-th data in the dataset.

Conditional probability. The conditional likelihood of each characteristic in the training data set is known as conditional probability. The formula:

$$P\left(X^{(m)} = a_{ml} \middle| Y = C_k\right) = \frac{\sum_{i=1}^{N} I\left(X_i{}^{(m)} = a_{ml}, y_i = C_k\right)}{\sum_{i=1}^{N} I(y_i = C_k)}$$
$$m = 1,2,\ldots,n, l = 1,2,\ldots,S_j, k = 1,2,\ldots,K \quad (2)$$

In the formula (3), The formula on the left represents the probability that X equals a_{jl}, when $y = C_k$. The denominator of the formula on the right represents the amount of data in the dataset equal to C_k. Molecular represents the number of data when y_i equals C_k, X equals a_{ml}.

Posterior probability. For the given instance $x_i = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})^T$, calculate the Posterior probability. The formula:

$$P(Y = C_k) \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} \middle| Y = C_k\right), k = 1,2,\ldots,K \quad (3)$$

In the formula (4), C_k is the data we request. $P(Y = C_k)$ is the probability of data with a value of C_k occurring in the dataset. $P\left(X^{(j)} = x^{(j)} \middle| Y = C_k\right)$ is the conditional probability of all C_k in the dataset.

Posterior probability. Determine the class of instance x based on the value of the maximum a posteriori estimation by calculating the maximum posterior probability. The formula:

$$y = \arg max\, P(Y = C_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = C_k) \quad (4)$$

2.2.2. BP model. The BP learning algorithm for multi-layer networks was first suggested by researchers led by Rumelhart and Mecelland in 1986. Here is one of its important steps: Signal forward propagation. Signal forward propagation. The input and output for calculating the forward transmission of signals are shown in Table 1.

Table 1. The input and output for calculating the forward transmission of signals.

Signal forward propagation	input	output
Enter the i-th node	x_i	x_i
The hidden layer's x-th node	$a_x = \sum_{x=1}^{m} y_{ix} x_i$	$b_x = f(a_x - \delta_x)$
The output layer's y-th node	$\beta_j = \sum_{h=1}^{m} w_{hj} b_h$	$\hat{y_j} = f(\beta_j - \theta_j)$

Table 1 shows the input and output of the i-th node as well as the connection weight between that node and the x-th node in the hidden layer as x_i and y_{ix}, respectively. The connection weights between the x-th nod and the j-th node are represented by the string w_{xj}. The weighted total of the results from the nodes in the preceding layer serves as the input to the hidden layer and output in layer nodes. a_x and β_j represent the input of the x-th node and the j-th node in the output layer. The output uses the Sigmoid Activation function, as shown in formula (6) below, in which x is the data to classify, to realize arbitrary nonlinear mapping from input to output. δ_x and θ_j represent the activation threshold. b_x and $\hat{y_j}$ represent the output of the x-th node in the hidden layer and the j-th node in the output layer.

$$f(x) = Sigmoid(x) = \frac{1}{1+e^{-x}} \quad (5)$$

2.2.3. Loss function. Choosing the appropriate loss function is of great significance to the model. The Mean squared error formula is continuous and differentiable, easy to calculate the gradient, and has good

DOI: 10.54254/2755-2721/37/20230467

mathematical properties. This enables the BP neural algorithm to use gradient descent and other optimization methods to minimize the Loss function so that the network can better fit the training data. Hence, in this study, the Mean squared error formula is used as the Loss function. The formula is below y is the predictive value and y^* is the standard value. MSE is the mean square error loss.

$$Loss = MSE(y, y^*) = \frac{\sum_{i=0}^{n}(y-y^*)^2}{n} \quad (6)$$

2.3. Implementation details

Some important aspects need to be emphasized in this study. Firstly, about hyperparameters: the Learning rate is 0.1. The training target accuracy is set to 0.01 and the maximum training frequency is 5000.

3. Result and discussion

In this chapter, the author first analyses the classification results of NB based on sepal length and width, as well as petal length and petal width, and proposes the applicability of the NB model. Afterward, the author analyses the loss function curves and accuracy curves of the BP model under different iterations and summarizes the characteristics of the loss and accuracy of the BP model under different iterations.

3.1. NB model

The classification outcome based on sepal length and sepal width is shown in Figure 3, and the classification outcome based on petal length and petal width is shown in Figure 4. The currency of the former is 79.8% while the latter is 96.8%. In contrast to predicting based on the length and width of sepals, using the length and width of petals is more accurate. This effect is caused by the increasing independence of petal length and width measurements. Therefore, NB models are more suitable for classifying data that are mutually opposed.

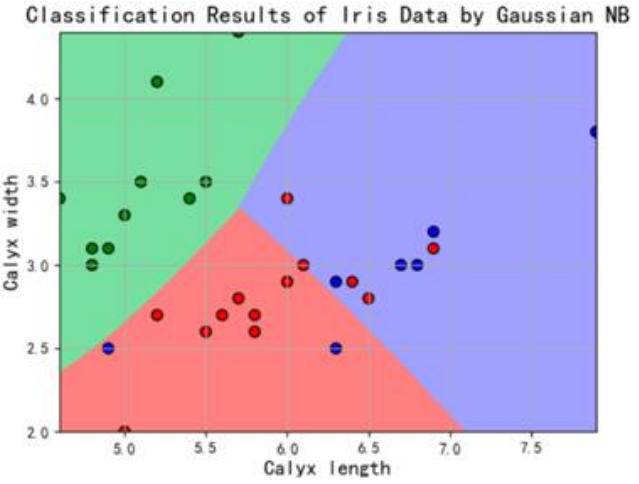

Figure 3. Results of classification using sepal width and length.

DOI: 10.54254/2755-2721/37/20230467

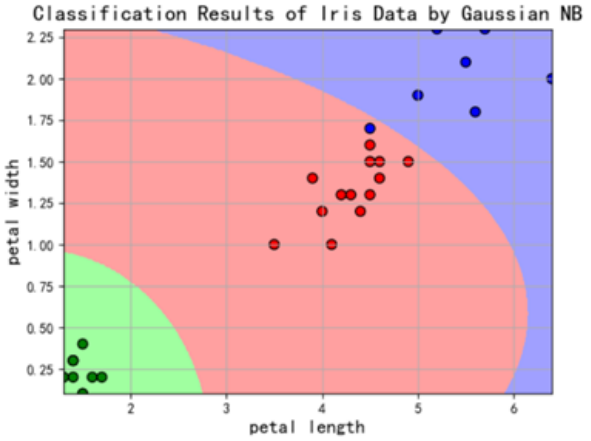

Figure 4. Results of classification using petal width and length.

3.2. BP model

Figures 5 and 6 show the loss function curve and accuracy curve for the BP model, respectively. Figures 5 and 6 show that when the number of iterations is low, the model's Loss function curve is steep and its accuracy is low. However, when the number of iterations rises, the model's accuracy curve dramatically rises and the Loss function curve sharply falls. The decreased pace of the Loss function curve and the increased speed of the model's accuracy curve slow down when the number of iterations hits 50. When iterations reach 700, the model's loss function curve practically stops declining, and the accuracy rate has always been 1. In summary, when the data is relatively independent, the NB model has a good classification function. When there are many iterations, the accuracy of the BP model is extremely high, even reaching 1. The NB model has a small amount of code, fast computation speed, and a complex BP model with high accuracy. Based on the traits of each of the two models, it is required to choose the best model for classification in real-world applications.

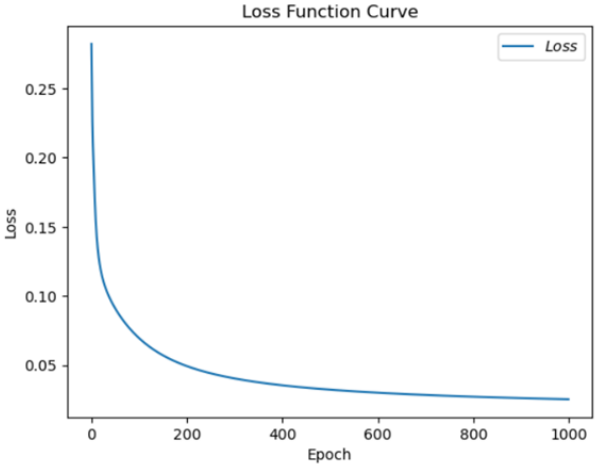

Figure 5. The loss function curve of the BP model.

DOI: 10.54254/2755-2721/37/20230467

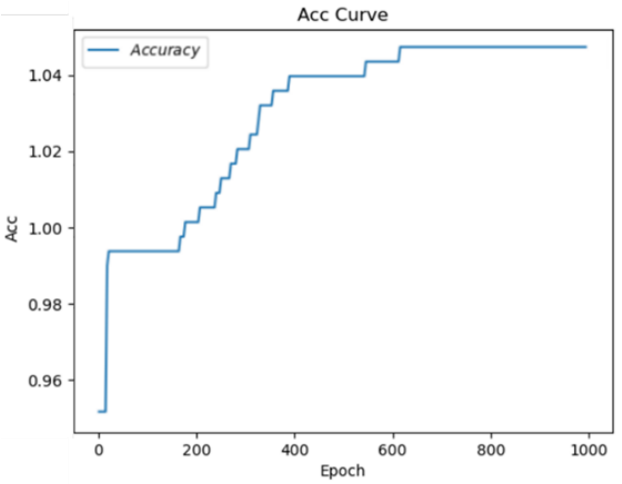

Figure 6. The accuracy curve of the BP model.

4. Conclusion

In this study, NB and BP are introduced to construct a model to classify Iris. Specifically, when using NB to classify the Iris dataset based on sepal length and sepal width, the accuracy is 79.8%, and the accuracy of classification based on petal length and petal width is 96.8%. Thus, the NB model is suitable for classifying relatively independent data. When using the BP model, the authors compared the accuracy and loss function of the model at different iterations. The model has good loss and accuracy when the number of iterations is between 300 and 700. The analysis found that compared to the traditional NB machine learning model, the use of a neural network algorithm for iris classification is more effective. Meanwhile, the neural network algorithm can mine the hidden patterns and information in the iris data. Future research will explore how to introduce a defense mechanism in the model to counter the attack of malicious data and improve the robustness of the model.

References

[1] Rumelhart D E Hinton G E Williams R J 1986 Representations by Back propagatingerrors Nature 323(6088): pp 533-536
[2] Mukherjee A Jain D K 2020 Back Propagation Neural Network Based Cluster Head Identification in MIMO Sensor Networks for Intelligent Transportation System IEEE pp 28524-28532
[3] Webb G I 2017 Naïve Bayes Encyclopedia of Machine Learning and Data Mining pp 1-2
[4] Kang C L 2020 Application of Neural Networks Learned by Mentors in Iris Species Recognition Xinzhou Demonstration School Report pp 17-21
[5] Pandey A Jain A 2017 Comparative Analysis of KNN Algorithm using Various Normalization Techniques I.J.Computer Network and Information Security pp 36-42
[6] Wu Y Y He J Ji Y M Huang G L Yao H C Zhang P Xu W Guo M J Li Y T 2019 Enhanced Classification Models for Iris Dataset Procedia Computer Science pp 946-954
[7] Rana D Jena S P Pradhan S K 2020 Performance Comparison of PCA and LDA with Linear Regression and Random Forest for IRIS Flower Classification PalArch's Journal of Archaeology of Egypt/Egyptology pp 2825-2830
[8] Hussain Z F Ibraheem H R 2020 A new model for iris data set classification based on linear support vector machine parameter's optimization pp 1079-1084
[9] Marques-Silva J Gerspacher T 2020 Explaining Naïve Bayes and Other Linear Classifiers with Polynomial Time and Delay Advance in Neural Information Processing System p 33
[10] Sang B 2020 Application of genetic algorithm and BP neural network in supply chain finance under information sharing Journal of Computational and Applied Mathematics 384
[11] Iris dataset https://archive.ics.uci.edu/dataset/53/iris