Comparison of Different Machine Learning Algorithms in Classification

Bingru Han*
School of Economics, Shanghai University, Shanghai, China

*Corresponding author e-mail: hbr1979@shu.edu.cn

Abstract. Classification, as a basic problem in the field of data analysis and machine learning, plays a more and more important role in human life. People are faced with huge amounts of information every day. How to classify information and how to extract useful information has gradually become a hot topic for scholars. There are many kinds of classifiers, such as neural network, support vector machine, decision tree, Bayesian classification algorithm and so on. This paper compares the classification results and accuracy of decision tree, support vector machine and naive Bayesian method by selecting data sets, and briefly describes its operation principle. The results show that the three machine learning classifiers perform well in dealing with the binary classification problem, but compared with the other two models, the decision tree model has higher accuracy.

Keywords: Machine Learning Classifier, Decision Tree, Support Vector Machine, Naive Bayes

1. Introduction
With the continuous development of science and technology, the popularity of the Internet is also higher and higher, people need to deal with a large amount of information every day, often these information are complex and chaotic, how to quickly classify them and then extract useful information by category, so that the distribution of information is more regular, is the task of classification[1]. Classification is a basic problem in the field of data science and machine learning. Its basic idea is to determine the label of input data through a function[2]. In recent years, more and more scholars turn their attention to the research of classification, and data-driven classifier learning has been a hot topic in recent years. There are many lots of machine learning classifiers, for example, neural network, decision tree, support vector machine, random forest, naive Bayes and so on[3].

Scholars at home and abroad have made great achievements in the research of machine learning classification model[4]. Liu Lilin et al.[5] compared the differences of CART decision tree, RF, MLPNN and CNN in the classification of steel plate defect data sets, and the results show that CNN has excellent classification ability. Yu Qiuyan et al.[6] reduced the dimension of the physical examination index of the data through principal component analysis, and then compared the effect of Bayesian network, support vector machine, random forest algorithm, neural network and decision tree
classifier on the classification and prediction of fatty liver. The results show that the decision tree classification model has good performance in the classification of fatty liver.

In this paper, iris dataset is taken as an example to compare the different performance of support vector machine, naive Bayes and decision tree classifiers in dealing with classification problems, which provides some ideas for further research on the application of machine learning classifier in classification problems. This paper is divided into four parts. The first part briefly describes the background of the problem and the previous research results in machine learning classifier. The second part describes the principles of several typical machine learning classifiers. The third part compares the performance results of different classifiers on the same data set and compares them. Finally, draw the conclusion.

2. Machine Learning Classifier

There are many machine learning classifiers that can handle classification tasks[7]. For example, classify the mail, infer the type of emotion and so on. This section mainly introduces decision tree model, classifier based on Naive Bayes algorithm and support vector machine model.

2.1. Decision Tree

Decision tree is a common machine learning method, which summarizes and classifies the attributes of the data in the training set, and finds an exact description and classification model for the attributes. Decision tree is named because its shape is like a tree and it can be used in decision-making. As a decision-making technology, decision tree uses a tree diagram to represent the expected value of each decision. By comparing the expected value, the decision scheme with the largest benefit and the smallest cost is selected. For the multi-stage and multi-level decision-making process, decision tree can express the relationship between decision-making and overall decision-making, and its mutual effect in a concise way because of its clear and vivid characteristics. The basic structure of decision tree includes node, decision node, event node and terminal node, which also includes probability branch and decision branch.

The model can classify the unpredictable data in the future. Decision tree classification is implemented by a variety of algorithms, such as ID3, C4.5, cart and so on[8], which are commonly used algorithms. In this chapter, ID3 algorithm is selected for classification, which is a classic decision tree classification algorithm, and constructs the decision tree from top to bottom. The partition standard of the algorithm is to get the information gain through the information entropy. Each partition selects the attribute with the highest information gain value as the standard, and repeats this process. The information entropy of the final sample is zero.

2.2. Naive Bayes

Naive Bayesian algorithm is a classification method based on Bayesian theorem and independent hypothesis of feature conditions. It is an extension and derivative of Bayesian classifier. It is called naive Bayes because it assumes that each input variable is independent. Although this condition is difficult to meet in actual data, naive Bayes is still very effective for most complex problems. It is one of the classifiers with better learning efficiency and classification effect.

Naive Bayes is a classification method based on Bayesian decision theory. It is a classification algorithm based on probability and statistics theory. Its core idea is: compare new data with known data set, select high probability results to make decision on new data[9]. Therefore, it is easy to train naive Bayes classifier. It does not need to use gradient descent algorithm and other complex algorithms to seek the optimal solution like other machine learning algorithms. Instead, it uses training data set to count the sample distribution of prior data in the case of selecting various features. Through a simple Bayesian formula, as shown in (1), it can get the maximum posterior probability Classification.
\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]  

By describing the causal relationship between variables, Bayesian classification algorithm combines prior knowledge with probability reasoning, which can train a classifier with better accuracy and help to optimize people's decision-making. As long as the prior probability of each class and the likelihood function of the feature are given, the naive Bayesian model can be trained, and the classification with the maximum posterior probability is the final classification. Naive Bayes algorithm is a relatively simple classifier based on Bayes theorem. It has a very good application in text analysis and has high stability, but it also has some shortcomings. Naive Bayes algorithm needs to meet the basic assumption that each feature is independent of each other, which is difficult to meet in real life, so naive Bayes algorithm is used to process data even when the correlation between data attributes is high, the effect of naive Bayes is not so ideal.

2.3. Support Vector Machine

\[ \min \frac{1}{2} \| \omega \|^2 \quad s.t. \ y_i \left( \omega \cdot X_i + b \right) \geq 1, i = 1, 2, ..., m \]  

Support vector machine model is a two classification model. Its basic idea is to find the best segmentation hyperplane in high-dimensional feature space, so that the distance between points in feature space and hyperplane is the maximum[10]. The learning strategy of SVM is the maximum interval[11]. The normal vector \( \omega \) and constant \( b \) make the hyperplane uniquely determined by defining the distance between the direction of hyperplane and its origin. Finally, it can be transformed into a solution of convex quadratic programming problem, as shown in (2).

SVM can be used to discuss classification and regression problems. This paper focuses on using SVM to deal with classification problems, so it is discrete label with a value of ±1. In order to calculate the duality problem of formula (2) by Lagrange multiplier method, it is more efficient and convenient to calculate. If the data is not linear separable, the kernel function is needed to simplify the problem. The technology adopts the idea of dimension raising and transforms the original low-dimensional data into higher dimension space by nonlinear mapping, which makes it satisfy the linear separable hypothesis in high-dimensional space. The selection of kernel function has a direct influence on the optimal solution of the problem. Generally, Gaussian kernel function is considered first according to experience. Support vector machine algorithm has a weak dependence on the number of samples, and has a high accuracy. Therefore, it has been widely used by scholars in practical research.

3. Results and Discussion

3.1. Data Selection

In this paper, iris data set is selected to verify the accuracy and efficiency of decision tree model, naive Bayes algorithm and support vector machine model in the classification problem. Iris data set is imported directly through sklearn package. This data set contains 150 records of three categories, 50 records of each category. Each record has four characteristics: calyx length, calyx width, petal length and petal width. Each characteristic is a continuous variable, which can be used to predict which of the three categories iris flower belongs to.

3.2. Classification Results

In the empirical part, this paper uses decision tree model, naive Bayes algorithm and support vector machine model to deal with iris data classification, and compares the results of three machine learning classifiers.
Because each characteristic variable in Iris data set is continuous variable, their distribution is unknown. When using naive Bayesian classifier, we need to assume the likelihood function of the sample. Here, we assume that the likelihood function of iris data set is Gaussian distribution. The accuracy of the three machine learning classifiers for iris dataset classification is shown in table 1.

**Table 1. Comparison of accuracy of three classifiers**

| Model    | SVM | Decision Tree | Naivie Bayes |
|----------|-----|---------------|--------------|
| accuracy | 0.86| 0.99          | 0.96         |

It can be seen from the above table that the accuracy of the models from high to low are: decision tree model, naive Bayesian model and support vector machine model. Generally speaking, these models have higher accuracy, but the accuracy of decision tree model is higher. Therefore, the decision tree model can be regarded as the optimal model to deal with the classification of iris dataset.

4. Conclusion

In this paper, we test the accuracy of decision tree model, naive Bayes model and support vector machine model in dealing with binary classification problems by selecting iris data sets. Through the above analysis, we can find that although compared with support vector machine model, naive Bayes model has higher accuracy, but it is not as high as decision tree model. In fact, support vector machine algorithm is different from naive Bayes algorithm.

Naive Bayes is a generation model, which needs to establish the distribution of the whole data, while support vector machine is a discriminant model, which only needs to establish the conditional probability distribution. So SVM only needs to see the data and separate them completely in space. Therefore, when the amount of data is small, SVM algorithm is greatly affected by noise and the classification effect is poor, while naive Bayes algorithm is better when it comes to small data sets. Naive Bayes model is the simplest Bayesian classifier, because it is subject to the basic constraint of all features independent, and can not show perfect accuracy. At present, many scholars have improved naive Bayes algorithm, trying to remove the constraint of feature independence. In future research, we hope to further explore the improved naive Bayes method Will the method perform better in the classification of data sets.

Generally speaking, the three models all have good performance in dealing with classification problems, so it will be meaningful to study how to combine several machine learning classifiers to deal with more classification problems in the future.

References

[1] Amasyali Kadir, and El-Gohary Nora."Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings." Renewable and Sustainable Energy Reviews 142,(2021):

[2] Palaiokostas Christos."Predicting for disease resistance in aquaculture species using machine learning models." Aquaculture Reports 20,(2021):

[3] Fermanian Adeline."Embedding and learning with signatures." Computational Statistics & Data Analysis 157,(2021):

[4] Wang Yalin, Chen ninren. Comparison of different machine learning algorithms in classification problems [J]. Heilongjiang science, 2021,12 (04): 16-18 + 22. (in Chinese)

[5] Liu Lilin, Tan Rong, Gao Xiang. Performance comparison of different machine learning models for steel plate defect classification [J]. Modern electronic technology, 2021,44 (01): 101-106. (in Chinese)

[6] Yu Quyan, Zhao Ying, Sun Ji, Shao Jianhua. Implementation and comparison of typical machine learning algorithms in fatty liver classification and prediction [J]. Journal of Mathematical Medicine, 2019,32 (01): 1-3.

[7] James Ming Chen."An Introduction to Machine Learning for Panel Data." International Advances in Economic Research .prepublish(2021):.
[8] Keting Ye. "Research on Personal Credit Risk Assessment based on Machine Learning Algorithm." World Scientific Research Journal 6.11(2020):

[9] Peng Ge. Survey of spam filtering based on Naive Bayes algorithm [J]. Computer knowledge and technology, 2020, 16(14): 244-245 + 247.

[10] Hongwei Zhu, Xuesong Wang, and Li Xiaolong. "An information security analysis method of Internet of things based on balanced double SVM." Journal of Intelligent & Fuzzy Systems 39.6(2020):

[11] Shakerin Farhad, and Gupta Gopal. "White-box Induction From SVM Models: Explainable AI with Logic Programming." Theory and Practice of Logic Programming 20.5(2020):