Stellar Classification by Machine Learning

Zhuliang Qi *

Civil Engineering, Chongqing University, Chongqing Municipality China, 400044

ABSTRACT: As an emerging subject with strong comprehensiveness, machine learning has made varying degrees of progress in various fields. In the field of astronomy, it has also been generally used, and there have been quantities of research using machine learning for data processing and model prediction. The paper has used three algorithms (Decision Tree, Random Forest and Support Vector Machine) to build prediction models to classify stars, galaxies, and quasars in the universe and make a comparison among three models. The results of the test have shown that the prediction accuracy of the Random Forest model reaches roughly 98 percent with a great computing efficiency, which performs the best.

1. INTRODUCTION

With the development of science and technology, many sky survey technology projects have been completed and put into operation, such as the Sloan Digital Sky Survey (SDSS) [1], making it easier to obtain data and information about various celestial bodies. At the same time, machine learning technology can now use various algorithms to train the massive data obtained from observation and collection for efficient classification [2]. It is of great significance not only to study the properties of various stars, but also to further explore the universe. Currently, there have been many achievements in academic research in this field. For example, in the research of star or galaxy classification based on Stacking ensemble learning, the base classifier model is established by using support vector machine, random forest and other algorithms. The gradient lifting tree is used as the meta classifier model. Finally, based on the classification accuracy of galaxies and other indicators, the classification results are compared with various machine learning algorithms. It is found that the accuracy is improved by nearly 10%. Another research is about the classification of star spectrum based on deep learning. Based on the convolutional neural network, the features of the star spectrum are extracted and used for classification. Compared with the traditional machine learning algorithm, the classification result has higher accuracy and robustness.

This test will use Python language, which has the advantages of simplicity, expansibility, and so on [3], and the library inside is sufficient. Based on the decision tree, random forest and support vector machine algorithm, the galaxies, stars and quasars in stars are classified, and the performance of different models is compared. This research can not only improve the low efficiency and accuracy of the original method in star classification, but also learn the performance of different training models on various objects through the classification results, which can also provide a certain reference for the selection of classifier model in an ensemble learning algorithm.

2. METHODOLOGY

2.1. Data

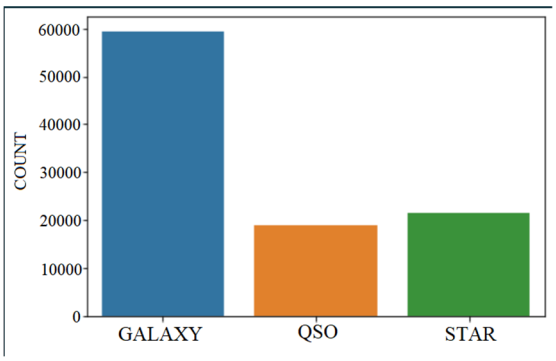

Figure 1 Sample size

The training data is spectral features observed by the Sloan Digital Sky Survey (SDSS). The data consists of 100,000 stars samples with 17 features. Among them, there are 59,445 examples of galaxies, 21,594 examples of stars and 18,691 examples of quasars, as shown in figure 1. In addition, each star sample contains 17 eigenvalues, among which the model focuses on the parameters measured by the SDSS photometry system in the U, G, R, I, and Z bands, as well as the alpha, Delta, and redshift of stars, as shown in figure 2.

*Corresponding author. Email: q1810641556@163.com

	alpha	delta	u	g	r	i	z	redshift
0	135.689107	32.494632	23.87882	22.27530	20.39501	19.16573	18.79371	0.634794
1	144.826101	31.274185	24.77759	22.83188	22.58444	21.16812	21.61427	0.779136
2	142.188790	35.582444	25.26307	22.66389	20.60976	19.34857	18.94827	0.644195
3	338.741038	-0.402828	22.13682	23.77656	21.61162	20.50454	19.25010	0.932346
4	345.282593	21.183866	19.43718	17.58028	16.49747	15.97711	15.54461	0.116123

Figure 2 Sample features

2.2. Data Preprocessing

As can be seen from the histogram of data distribution in the figure above, there is a large imbalance in the data. That is, the large difference in the sample numbers of galaxies, quasars, and stars will make the training model unable to effectively learn the decision boundary, resulting in a large deviation [4]. In order to solve this problem, the data will be oversampled by using the SMOTE function in the imbalanced-learn library of Python.

The working principle of the SMOTE function is to randomly select a minority of sample points, then find the k points that are closest to the selected points, and connect them with the sample points into line segments, so as to create more comprehensive samples for the minority data [5]. The samples created by the comprehensive oversampling method are relatively close to the existing few samples in the feature space, which enables the classifier to establish a large decision region containing a few nearby points. By SMOTE function, the minority data are oversampled and acquire new data. As shown in figure 3, the number of all kinds of star samples is up to 59445.

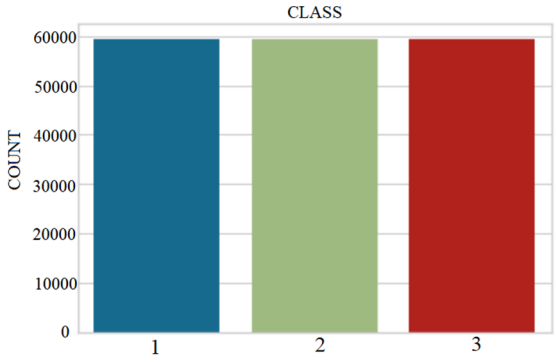

Figure 3 Number of each category after pretreatment

Considering that the subsequent operation will apply to the Support Vector Machine algorithm, the data should be normalized preprocessed to accelerate the accuracy and efficiency of model training [6]. The Standardscaler function in sklearn library will be used here. Its working principle is to de-mean and normalize each feature dimension in the data, so that the processed data can conform to the standard normal distribution. The transformation function is shown as follows:

$$x^* = \frac{x - \mu}{\sigma} \quad (1)$$

After normalization, the data sets are acquired as shown in figure 4:

	0	1	2	3	4	5	6	7
0	-0.426906	0.393940	0.084453	0.071610	0.372057	-0.068467	-0.005844	-0.084547
1	-0.333473	0.332480	0.118683	0.092822	1.585361	1.071683	0.101716	0.080121
2	-0.360442	0.549438	0.137173	0.086420	0.491064	0.035641	0.000050	-0.073822
3	1.649453	-1.262730	0.018108	0.128827	1.046259	0.693844	0.011560	0.254905
4	1.716345	-0.175655	-0.084710	-0.107331	-1.787821	-1.884049	-0.129747	-0.676254

Figure 4 Normalization

Galaxies, stars and quasars will be replaced with numbers 1, 2 and 3 respectively to facilitate subsequent model training operations. Finally, the data set is divided for model training and detection with a ratio of 0.3. (70% training set, 30% test set).

2.3. Model Introduction

2.3.1. Decision tree

As a common machine learning algorithm, decision tree is usually used for classification. As its name suggests, a decision tree simulates the basic structure of a tree for classification, including a root node, some internal nodes and leaves. Among them, each leaf node corresponds to a predicted result, while the sample data in each node will be allocated to the corresponding child nodes according to the results of attribute tests [7]. The root node contains the complete data set of the sample, and the path from the root node to each leaf node corresponds to the sequence of tests. The purpose of decision tree is to generate a model with strong generalization ability.

In this decision tree model training, the information entropy is used to measure sample purity, that is, to select

the optimal division. The definition formula of information entropy is:

$$Ent(D) = -\sum_{k=1}^{|y|} p_k log_2 p_k \quad (2)$$

Among them, the lower the information entropy value is, the higher the purity of partition is improved.

2.3.2. Random forests

Random Forest is an extended variant of Bagging. It builds the Bagging based on the decision tree and introduces the shorthand selection of attributes in the training process of the decision tree [8]. In other words, the decision tree compares the purity of differentiation according to the information entropy, gain rate and other indicators when it chooses the attributes, so as to select an optimal attribute. In Random Forest, a subset containing K attributes is randomly selected from the attribute set of each node of the base decision tree. Then an optimal attribute is selected from this subset for division. In the development and application of Random Forest, its advantages of simplicity, easy implementation and low computational cost lead to strong performance in realistic tasks.

2.3.3. Support Vector Machine

The Support Vector Machine (SVM) is a binary classification model that seeks linear classifiers with the greatest distance from sample boundaries in the feature space [7]. The basic functional form of its partition plane is shown as follows:

$$\omega^T x + b = 0 \quad (3)$$

While ω represents the direction of the plane, b represents the displacement term, which is the distance between the plane and the origin. The so-called "interval" refers to the sum of the distance between the sample points closest to the boundary plane and the plane, which can be expressed as:

$$\gamma = \frac{2}{\|\omega\|} \quad (4)$$

The working principle of SVM is interval maximization, which can be specifically interpreted as solving the problem of convex quadratic programming. For a linearly separable data set, countless hyperplanes can be found to divide the samples, but the hyperplane with the largest distance to the nearest point of the sample is unique [9]. Hence, the support vector machine algorithm seeks to find this hyperplane to achieve optimal classification. For the nonlinear sample distribution, support vector machine also includes kernel skills. Through the combination of multiple functions to form nonlinear boundary, it can also be transformed into a linear classification problem in a higher dimensional feature space.

3. RESULT AND DISCUSSION

3.1. Decision tree

This training will use the DecisionTreeClassifier function of sklearn. The classification report is shown in figure 5.

	precision	recall	f1-score	support
1	0.95	0.95	0.95	17737
2	0.99	0.99	0.99	17847
3	0.96	0.96	0.96	17917
accuracy			0.97	53501
macro avg	0.97	0.97	0.97	53501
weighted avg	0.97	0.97	0.97	53501

Figure 5 Classification report

It can be seen that the accuracy rate of decision tree model for galaxies is 95%, the recall rate is 95%, and the F1 score is 95%, while the accuracy rate, recall rate and F1 score of stars are roughly 99%; For the classification of quasars, the accuracy, recall and F1 scores are all 96%. The overall accuracy of the model is 97%, and the model training takes 0.6 seconds. The model of the decision tree has high accuracy, the fastest operation efficiency among the three models, and the best performance in star classification. In addition, sample training results can be evaluated by the confusion matrix and ROC curve. Among them, the confusion matrix indicates the details of the classification by counting the number of correct and wrong classification for each category. The ROC curve can perfectly reflect the performance of the model through AUC parameters [10]. The larger the AUC value is, the better the classification is.

As shown in figure 6, from the confusion matrix, it indicates the number of correct classifications and misjudgments in the classification of different categories by the decision tree model. The ROC curve, as shown in figure 7, shows that the decision tree model has good performance in the classification of the three categories. Among them, the classification of stars is the best, that is, AUC=1.00

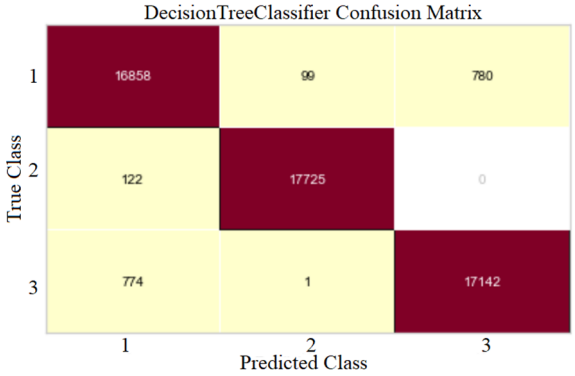

Figure 6 Confusion matrix

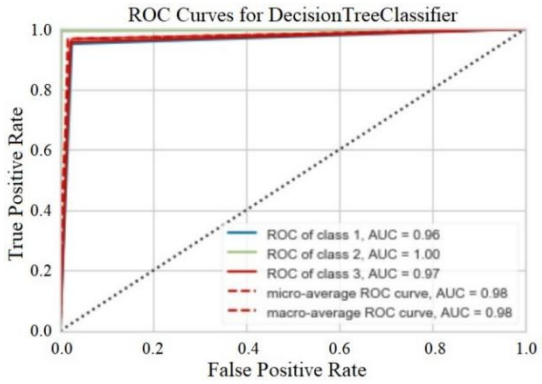

Figure 7 ROC curves

3.2. Random Forest

This training will use the Random Forest Classifier function in the sklearn library, and the training result is shown in figure 8:

	precision	recall	f1-score	support
1	0.97	0.98	0.97	17737
2	1.00	1.00	1.00	17847
3	0.98	0.97	0.98	17917
accuracy			0.98	53501
macro avg	0.98	0.98	0.98	53501
weighted avg	0.98	0.98	0.98	53501

Figure 8 Classification report

It can be found that the accuracy rate of random forest model for galaxies is 97%, the recall rate is 98% and the F1 score is 97%. For stars, the precision rate, recall rate, F1 score are 100%. For quasars classification, the accuracy rate is 98%, the recall rate is 97%, and the F1 score is 98%. The overall accuracy of the model is 98%, and the model training takes 26.2 seconds. It has the highest precision rate among the three models with high operating efficiency, and it performs the best.

The confusion matrix shown in figure 9 indicates the details of the classification by the random forest model. In the classification of stars, 17,844 are correctly classified while only 3 are incorrectly classified.

As can be seen from the ROC curve in figure 10, the random forest model performs well in the classification of three different types of stars, with AUC values of 1.00.

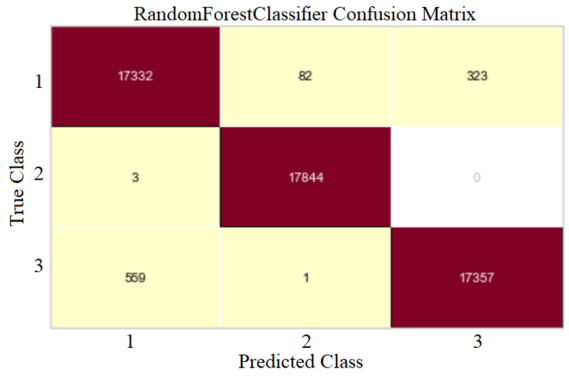

Figure 9 Confusion matrix

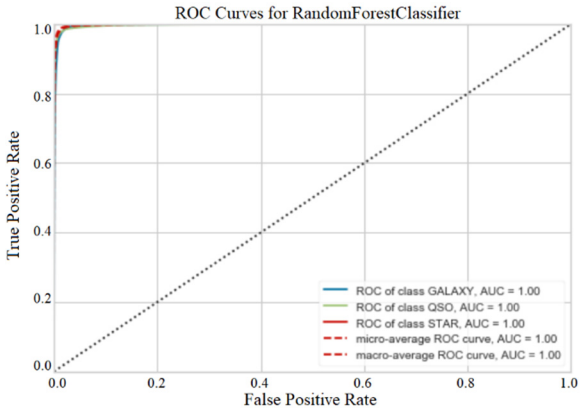

Figure 10 ROC curves

3.3. Support Vector Machine

SVM function in sklearn library will be used in this training, and the training results are shown in figure 11:

	precision	recall	f1-score	support
1	0.95	0.95	0.95	17737
2	0.97	1.00	0.98	17847
3	0.98	0.95	0.96	17917
accuracy			0.97	53501
macro avg	0.97	0.97	0.97	53501
weighted avg	0.97	0.97	0.97	53501

Figure 11 Classification report

It can be seen that the accuracy rate, recall rate and F1 score of the support vector machine model for galaxies are 95%. For stars, the accuracy rate is 97%, the recall rate is 100%, and the F1 score is 100%. For the classification of quasars, the accuracy rate is 98%, the recall rate is 95%, and the F1 score is 96%. The overall accuracy is 97%, and the total training time of the model is 6 minutes and 46.9 seconds. The accuracy rate of the model is high, but the training time is the longest with the lowest operational efficiency. In the confusion matrix shown in figure 12, compared with quasars and galaxies, the support vector machine model performs the best in the classification of stars. Among the 17,847 samples of stars, 17,837 are correctly classified, while only 10 are incorrectly classified.

As can be seen from the ROC curve in figure 13, the AUC value reaches roughly 1.00 with high accuracy in the classification of stars, while for quasars and galaxies, the AUC values are 0.99 and 0.96, respectively.

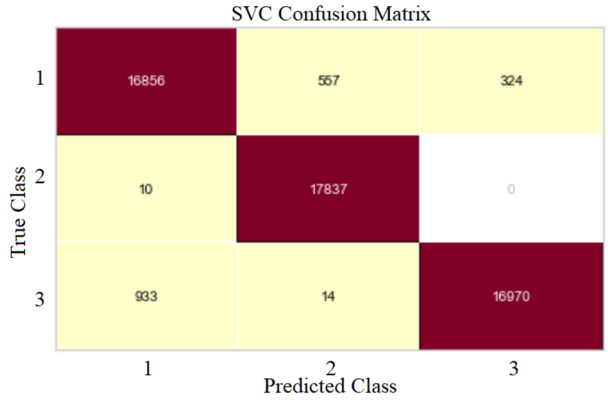

Figure 12 Confusion matrix

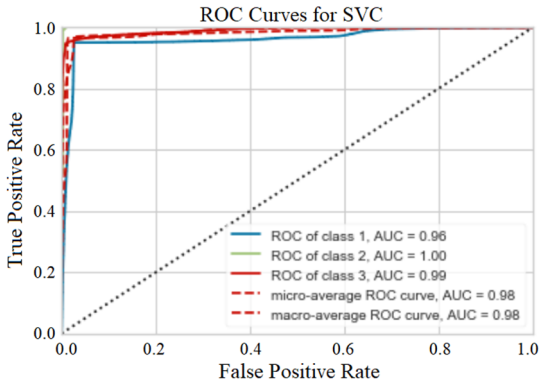

Figure 13 ROC curves

4. CONCLUSION

Based on the spectral data obtained from the Sloan Digital Sky Survey, this study trained the model with decision tree, random forest and support vector machine algorithms in machine learning, and classified the galaxies, stars and quasars in the universe. From the training results of the three models, it can be seen that the random forest algorithm has the best performance in the dataset, which not only has the highest accuracy rate of 98%, but also has a high computing efficiency. Compared with it, decision tree and support vector machine algoriths have a high accuracy rate of 97%, but support vector machine algorithm needs the longest time for calculation, whose efficiency is the lowest. Besides, all three algorithms performed well in the classification of galaxies, stars and quasars. As for the best performance in the star classification, it may be caused by the obvious difference between stellar and the other two stars in nature, resulting in the large differences in data. For the training, there are also some shortcomings that need to be improved. For example, in data processing, the combination of under-sampling and over-sampling can better avoid over-fitting and improve the training accuracy.

REFERENCES

1. YORK D G, ADELMAN J, ANDERSON JR J E, et al. The sloan digital sky survey: technical summary[J]. The Astronomical Journal, 2000, 120(3); 1579
2. Zhou Jie, Zhu Jianwen. Research on machine learning classification problem and algorithm [J] Software, 2019, 40 (7): 205-208
3. Liu Shifang. Application of Python in Artificial Intelligence. Industrial Technology Innovation, 2019, 25
4. Lin Zhiyong. The Effect of Data Imbalance and Others on SVM Classifiers—Experimental Study. Journal of Guangdong Polytechnic Normal University, 2008, 06
5. Wang Yao, Zheng lie, A New Tentative SMOTE Algorithm Based on Clustering. Journal of Chongqing University of Technology (NATURAL SCIENCE), 2021
6. Zhang Ge. Research on data preprocessing in course recommendation prediction Model, China New Telecommunications, 2019, 21(19)
7. Zhou Zhihua. Machine learning Beijing: Tsinghua University Press, 2016
8. Zhao Mingmei, Jin Yangyang, Wang Yujia, Zeng Mengjia. Application Research of Random Forest Algorithm in Decision Making. Computer & Network, 2021, 22
9. Wang Xia, Dong Yongquan, Yu Qiao, GENG Na. Review of Structural Support Vector Machines. Computer Engineering and Applications, 2020, 17
10. Shi Haosu. Comparative research of the ROC curve drawing based on case and MATLAB, Electronic Design Engineering, 2010, 9