DenseNet-random forest model based galaxy classification

Junyi Hu

Information Management and Information Systems, Beijing Jiaotong University, Weihai, 264200

22711012@bjtu.edu.cn

Abstract. Finding an efficient and accurate adaptive method that can automatically classify galaxies has become an industry consensus. However, most of the current studies on galaxy classification use a single model for direct output, without considering the combination with other models to output more satisfactory prediction results. Through convolutional neural network and classifier, this study studied the possibility of applying the deep learning model to the Galaxy 10 DECals dataset classification, and proposed DenseNet-Random Forest model through comparative analysis. By adjusting and training DenseNet-121 with appropriate hyperparameters, the input tensor is transferred to the basic model through the creation of a shape input layer, where GlobalAveragePooling2D is added to perform an average pooling operation on each feature map, reducing the spatial dimension of each feature map to 1. During the process, a complete connection layer with 64 neurons was added using the ReLU activation function, and a Dropout layer was added to randomly discard 20% of the neurons during training to prevent overfitting. In addition, ReLU Activation function with 32 full connection layers of neurons and softmax Activation function with 10 output layers of neurons are added. By acquiring the feature vector of the training model and the real label of the verification set, assign x and y values respectively, and import them into the Random forest classifier model. The experimental results demonstrated the model ultimately achieved a prediction accuracy of 68% when processing the Galaxy 10 DECals dataset, and achieved nearly 30% improvement in Precision, Recall, and F1 scores.

Keywords: DenseNet, random forest, galaxy classification.

1. Introduction

A galaxy represents a vast assemblage of celestial entitles, consisting of stars, gas, dust and dark matter, which is the largest celestial structure in the universe, consisting of tens of billions of stars and other celestial bodies. Exhibiting diverse morphologies and configurations, galaxies play a pivotal role in the cosmos. According to the shape and structure of galaxies, galaxies are mainly divided into spiral galaxies, elliptical galaxies, irregular galaxy, dwarf elliptical galaxies and other types of galaxies [1]. In the observable universe, the total number of galaxies may reach over 100 billion. Therefore, understanding the characteristics and types of galaxies is a crucial step for humans to explore the mysteries of the universe.

Astronomical observations of galaxies are typically conducted utilizing various radio telescope and spectrometers, and obtain their pictures through a series of analytical processes. The classification of galaxies represents a fundamental and indispensable pursuit within the field of astronomy. Initially, astronomers classified galaxies manually based on their different physical characteristics. However, due to the complex morphology and characteristics of galaxies, as well as their massive data, there are high requirements for professional knowledge and relevant experience, resulting in the disadvantages of low efficiency and limited accuracy in manual classification. In addition, influenced by individual subjectivity and preferences, manual classification can lead to a certain error rate in the classification results. Therefore, an adaptive method that can automatically classify galaxies to improve the efficiency and accuracy of galaxy classification has become an industry consensus.

Deep learning, as a branch of machine learning, has achieved tremendous development in recent times, which has been extensively applied to computer vision, natural language processing, augmented reality and virtual reality and other fields. In particular, deep learning methods have been introduced into the classification of galaxies. For example, the Fang et al. proposed a novel method different from traditional data augmentation - adaptive polar transformation, which considers the rotation invariance of the Convolutional Neural Network (CNN) model in the preprocessing step and transforms the rotation invariance problem into a translation invariance problem for galaxy classification [2]. Tarsitano et al. used isophotometric fitting to analyze the main features of two-dimensional light distribution in galaxy images, and used the integration and Principal Component Analysis (PCA) decomposition of XGBoost and random forest models as a feature engineering method to classify galaxies [3]. The Ghadekar et al. proposed a ConvNet galaxy architecture for classifying galaxies by constructing a Deep CNN framework and utilizing different features [4]. Radhamani et al. used box counting algorithm to calculate fractal dimension as the main feature of different types of galaxies, and used LeNet-5 network model to classify galaxy images according to their morphological characteristics [5]. Iprijanovi et al. used DeepAdversaries to test and improve the robustness of the deep learning model of galaxy morphology classification [6]. The Berna et al. used Fast Fourier Transform as a preprocessing step and convolutional neural networks to classify galaxies [7].

However, the proposed deep learning methods mentioned above in most recent studies on the identification and classification of galaxies is based on the direct output of a single convolutional neural network, and does not consider trying to combine with other models to output the prediction results based on the more satisfactory performance. To address this issue, this paper is based on the Galaxy 10 DECals dataset, with a focus on exploring the processing effect of convolutional neural networks and classifiers on galaxy classification. It combines the densenet network with different classifiers for output, and explores the optimal models corresponding to the densenet network through comparative analysis. The results show that the densenet network combined with the random forest classifier achieves 68% classification accuracy.

2. Method

2.1. Dataset description and preprocessing

The Galaxy 10 DECals dataset is a collection of spectral data for 10 various objects used by astronomers in their astronomy research [8]. It was born from the DECals project, which aims to provide astronomers with a complete vision of the sky, thereby facilitating the exploration of large-scale structure in the universe. The Galaxy10 dataset classification labels are sourced from Galaxy Zoo, while the corresponding images are procured from DESI Legacy Imaging Surveys. The Galaxy 10 DECals is a galaxy morphology image dataset categorized into 10 categories as shown in Figure 1, containing 17,736 galaxies (g, r, and z bands) color images of 256×256 pixels.

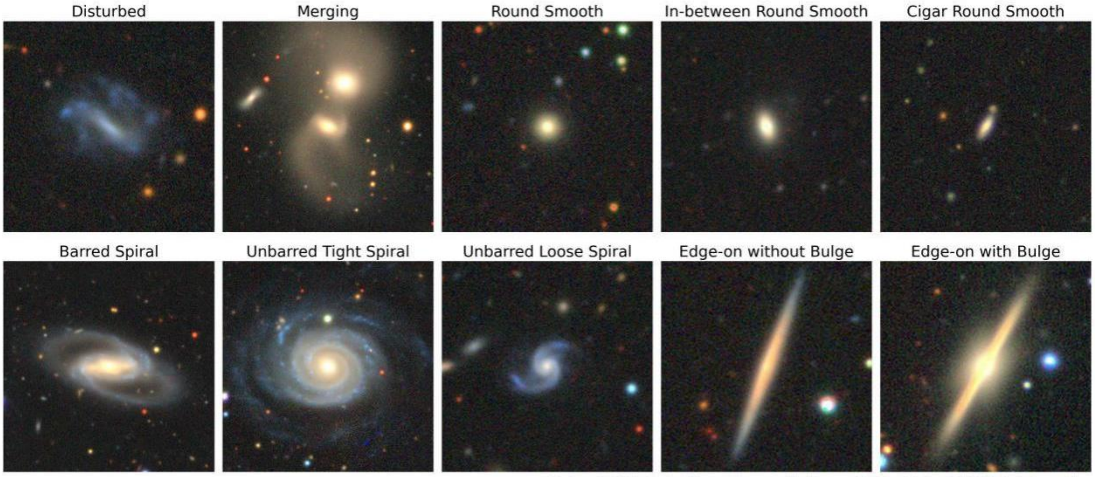

Figure 1. The sample images of the ten categories of galaxied in the collected dataset.

To preprocess the data, this study employs ImageDataGenerator. In the realm of deep learning, the issue of sample balance holds significant importance. Imbalanced sample sizes across categories can bias the model and affect its accuracy. Due to the large variance in the number of galaxies in each class, the sample balance problem must be considered in galaxy classification tasks. In the Galaxy 10 DECals dataset, the sample size of some categories is much larger than that of other categories. Owing to this category imbalance, the model may over-focus on the category with a large sample size and ignore other categories, which will have an impact on the training of the model. ImageDataGenerator is a widely used data augmentation technique that can increase the diversity of datasets without increasing the sample size, and improve the robustness, scalability, and generalization capabilities of classification models. The aspect ratio of galaxy images will have an impact on the identification of galaxy types due to their unique features. ImageDataGenerator performs data standardization on the image, processes data enhancement data on the color channel, and maintains the consistency of the image scale by moving and rotating. ImageDataGenerator converts the labels into the form required by the model according to different representations of the categories, and enhances the classification ability of the model. The data augmentation for this study adopts random horizontal flip and random rotation of 0.2.

2.2. CNN-RF based galaxy classification

CNN has emerged as a prominent deep learning model extensively employed in computer vision tasks and image recognition. Among them, the convolutional layer is used to extract features from the input data. The pooling layer is used to reduce the size of the feature map and extract the main features. The alternating appearance of convolutional layer and pooling layer in CNN constitutes a deep network structure. Dense Connection Network (DenseNet) shown in Figure 2 is a deep convolutional neural network architecture that promotes the flow of information and the reuse of features through dense connections, aiming to solve the problems of gradient disappearance and insufficient parameter utilization in deep network training. Unlike traditional convolutional neural networks, the output of each layer in DenseNet is not only connected to the input of the next layer, but also connected to the input of all subsequent layers through the combination of dense blocks and transition layers.

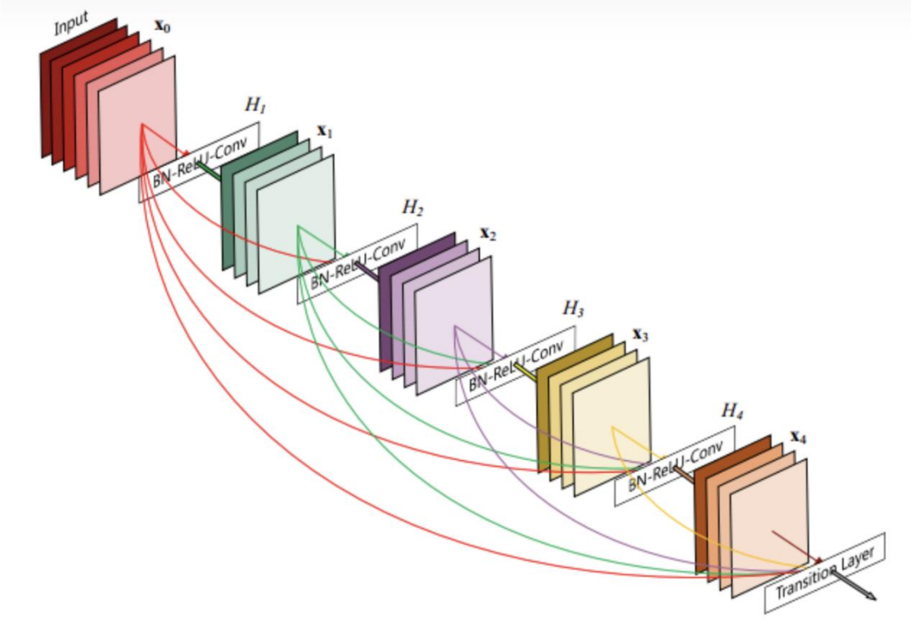

Figure 2. The basic framework of the DenseNet [9].

Random Forest (RF) is an ensemble learning method that makes a final prediction by voting or averaging the prediction results of multiple decision trees for classification and regression tasks. RF can reduce the risk of overfitting and improve the accuracy and robustness of the model by integrating multiple decision trees. Furthermore, RF can handle high-dimensional data with a large number of features without feature selection or dimensionality reduction.

The version of DenseNet (DenseNet-121) that employed in this study has 121 layers and was developed by Huang et al. in 2017 [9]. The neural network model adopts the initial DenseNet design with proper hyperparameter tuning and training. Passing the input tensor to the base model by creating an input layer of shape (128, 128, 3). Then adding a GlobalAveragePooling2D() to perform an average pooling operation on each feature map, reducing the spatial dimension of each feature map to 1. Initially, reducing the spatial dimension of each feature map to 1. Additionally, adding a full connection layer with 64 neurons using ReLU activation function and adding a Dropout layer to randomly discard 20% of neurons during training to prevent overfitting. Furthermore, adding a ReLU activation function with 32 full connection layers of neurons and a softmax activation function with 10 output layers of neurons. Changing the output of the model to a probability distribution containing 10 elements, which can be used for galaxy classification tasks. Sequentially, creating a new model using the specified input and output tensors. By acquiring the feature vector of the trained model and the true label of the verification set, assign values to X and y respectively, and import them into the random forest classifier model.

2.3. Implementation details

This study used Tensorflow as the deep learning framework since it has been chosen by many studies [10, 11]. Additionally, the GPU A100 was chose as the hardware accelerator to run in the environment of Colab. During the training process, Adam is used as the optimizer and categorical_ Crossentropy is used as the loss function. To avoid overfitting, the callback function of Early Stopping is used. Meanwhile, set the Learning rate to 0.0005, Batch size to 64, and Epochs to 10. Accuracy is the evaluation indicator for this learning session. The structure and design of DenseNet-Random forest model are shown in Figure 3.

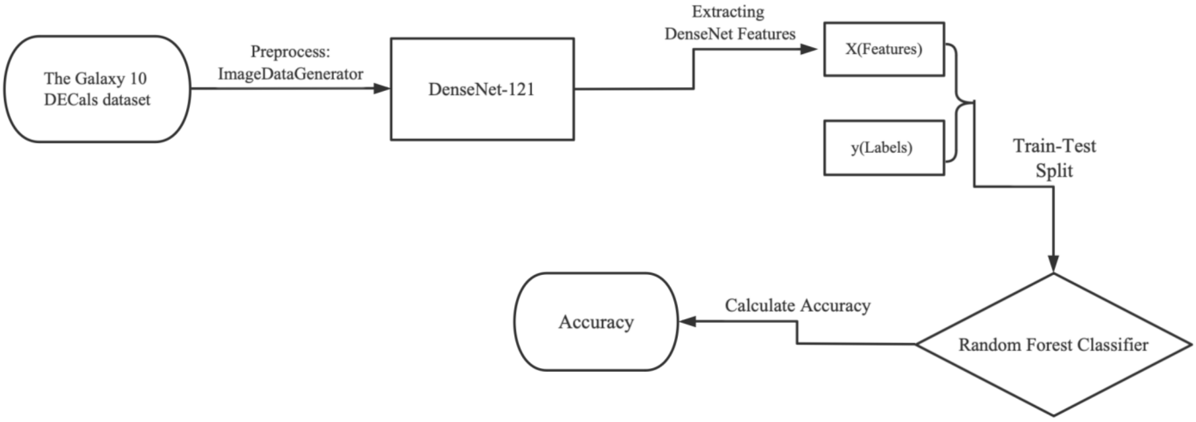

Figure 3. The structure and design of DenseNet-Random forest model.

3. Results and discussion

By adjusting epoch, batch size, a comparison was implemented between the single DenseNet model and the DenseNet-RF model. The experimental results demonstrated that the DenseNet -RF model outperformed the single DenseNet model, manifesting an performance improvement in accuracy from 37% to 68% shown in Table 1. Furthermore, it has obtained significant optimization and ascension in various aspects such as Precision, Recall, F1 score, etc.

Table 1. The performance of densenet combined with random forest model in the Galaxy 10 Decals dataset.

Model	Performance			
	Accuracy	Precision (weighted avg)	Recall (weighted avg)	F1-score (weighted avg)
Single DenseNet	0.37	0.49	0.37	0.34
DenseNet+Random Forest	0.68	0.68	0.68	0.68

DenseNet is a well-established deep convolutional neural network architecture. Its densely connected design has unique advantages in feature reuse and parameter sharing, and it performs well in processing high-dimensional features and complicated features. DenseNet adopts a dense connection structure, which forms a dense information flow by directly connecting the output of the previous layer with the input of the subsequent layer. Random forest is a powerful ensemble learning model. This dense connection design makes the network have a stronger feature reuse ability, which aids in alleviating the issue of gradient disappearance, strengthen information transmission, and improve the effectiveness and expressiveness of the network. Random forests can capture the diversity and complexity in datasets and reduce over-reliance on specific features. By combining the prediction results of multiple decision trees, it can effectively reduce the risk of overfitting of a single decision tree, providing accurate prediction results. The DenseNet-RF model combines the advantages of the two models in a reasonable manner, which helps to overcome the limitations of a single model and improve the accuracy and robustness of predictions. The outcomes presented in the tabulated results allude to the substantial enhancements realized by the DenseNet-RF model, emphasizing its efficacy in improving overall model performance and elevating the precision of prediction outcomes.

4. Conclusion

This study investigated the possibility of using deep learning model on the classification of Galaxy 10 DECals dataset by convolutional neural network and classifier, and proposed a DenseNet-RF model through comparative analysis. After experiments, the experimental results demonstrated that under various and appropriate hyperparameter settings, the model significantly improves the accuracy of DenseNet when dealing with complex data sets, and effectively reduces the risk of overfitting. The model ultimately achieved a prediction accuracy of 68% when processing the Galaxy 10 DECals dataset, and achieved nearly 30% improvements in Precision, Recall, and F1 score. In the future, a combination of multiple neural network architectures and traditional classifiers such as vector machines will be considered to explore whether the combination of models can achieve better performance on different tasks and data sets.

References

[1] NASA 2023 Types NASA https://universe.nasa.gov/galaxies/types/.

[2] Fang G et al 2023 Automatic Classification of Galaxy Morphology: A Rotationally-invariant Supervised Machine-learning Method Based on the Unsupervised Machine-learning Data Set. The Astronomical Journal 165(2) 35.

[3] Tarsitano F et al 2022 Image feature extraction and galaxy classification: a novel and efficient approach with automated machine learning. Monthly Notices of the Royal Astronomical Society 511(3) pp 3330-3338.

[4] Ghadekar P et al 2023 Galaxy Classification Using Deep Learning In Advancements in Smart Computing and Information Security: First International Conference ASCIS 2022 Rajkot India November 24–26 Revised Selected Papers, Part I pp 3-13 Cham: Springer Nature Switzerland.

[5] Radhamani P S et al 2022 An Effective Galaxy Classification Using Fractal Analysis and Neural Network In 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) pp 19-24 IEEE.

[6] Ćiprijanović A et al 2022 DeepAdversaries: examining the robustness of deep learning models for galaxy morphology classification Machine Learning: Science and Technology 3(3) 035007.

[7] Bernal S S et al 2023 Galaxy Classification Upgraded: Boosting Machine Learning with the Fourier Transform Available at SSRN 4428890.

[8] Henrysky 2022 (n.d.) Henrysky/Galaxy10: A CIFAR10-like galaxy image dataset GitHub https://github.com/henrysky/Galaxy10.

[9] Huang G 2018 Densely connected Convolutional Networks arXiv.org https://doi.org/10.48550/arXiv.1608.06993.

[10] Yu Q Yang Y Lin Z et al 2020 Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV China Communications 17(3): 46-57.

[11] Kazakov O D Mikheenko O V 2020 Transfer learning and domain adaptation based on modeling of socio-economic systems Бизнес-информатика 14 7-20.