Publications of the Astronomical Society of the Pacific, 132:024504 (13pp), 2020 February https://doi.org/10.1088/1538-3873/ab5ed7

Classification for Unrecognized Spectra in LAMOST DR6 Using Generalization of Convolutional Neural Networks

Zi-Peng Zheng[1], Bo Qiu[1], A-Li Luo[2,3,4], and Yin-Bi Li[2]
[1] School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, 300401, People's Republic of China; qiubo@hebut.edu.cn
[2] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, People's Republic of China lal@nao.cas.cn
[3] University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China
[4] Department of Physics and Astronomy, University of Delaware, Newark, DE 19716, USA
Received 2019 August 12; accepted 2019 December 4; published 2020 January 13

Abstract

Commonly used template classification for celestial spectra always fails dealing with low signal-to-noise ratio (S/N) spectra, which are very numerous in spectroscopic surveys. In the sixth data release of Large sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST DR6 V1), more than 0.7 million bad quality data were refused to classify by LAMOST pipeline and archived as "UNKNOWN." To recognize as many objects with low S/N spectra as possible in the "UNKNOWN" data set, one-dimensional convolutional neural network (CNN) based classifier was adapted from the widely used two-dimensional CNN. In this work, two CNN based classifier were applied, a classifier for distinguishing galaxy, QSO and star, and a classifier for discriminating subtypes of stars. To solve the problem caused by imbalanced training samples among different classes for the stellar classifier, a semi supervised learning algorithm by two CNNs and Spectral Generative Adversarial Network (SGAN) was introduced to produce artificial spectra for the minority O type. The SGAN solution is better than over-sampling in solving overfitting caused by imbalanced training set. The trained CNN classifiers were applied to classify "UNKNOWN" spectra into candidates of galaxies, QSOs, and stars. and further classify star candidates into spectral subclasses of O to M. Each spectra can be recognized to a spectral type with a probability by CNN algorithm, and 101,082 stellar spectra were remained with the probability larger than 99%, making up a supplemental star catalog of LAMOST DR6, which includes 294 O, 2 850 B, 269 A, 6 431 F, 626 G, 60 527 K, and 30 085 M types. To verify the catalog, the distances to corresponding templates from recognized spectra in each class were also checked comparing with known spectra. In addition, 200 O type stars were manually confirmed from 294 automatically identified O type stars in the catalog, because O type spectra have weak features and easily to be confused with no signal spectra. The classification result as a part of this work are available at http://paperdata.china-vo.org/Classification_SGAN/result.zip.

Key words: catalogs – methods: data analysis – techniques: spectroscopic

Online material: color figures

1. Introduction

In recent years, more and more large-scale spectral surveys have been performing or completed, yielding unprecedented numbers of spectra of various celestial objects, which promote the vigorous development of research in various fields of astronomy. The celestial spectra contains physical properties, chemical compositions, and kinematics information of the objects, which could be derived through spectral analysis. Generally, the first step of spectral analysis is to classify and identify spectra, which is the basis for further study. For any large spectral database, automated classification methods are required, and many previous works (LaSala 1994; Bailer-Jones 1997; Carricajo et al. 2004; Bailer-Jones et al. 2008) focused on distinguishing large quantity spectra of different celestial type. However, these methods always fails when dealing with low signal-to-noise ratio (S/N) data.

The most commonly used spectral classification method is to compare observed spectra with theoretical or empirical spectra templates, namely cross-match, which is also adopted by the LAMOST pipeline (Luo et al. 2015). The pipeline cross-matches each observed spectrum with a set of templates to calculate chi-square values, and select the smallest value representing the class that the object belongs to. However for a low S/N spectrum, the confidence of the chi-square value of its best-fitted template is sometimes too low, so that the pipeline cannot judge its class and only has to label it as "UNKNOWN" (Guo et al. 2019). In this paper, we try to solve the classification problem of low S/N spectra with the help of data-driven method.

Artificial neural network (ANN) has long been applied in spectral classification, for example, Weaver & Torres-Dodgen (1995) used ANN to classify the near-infrared spectra of

A-type stars, Folkes & Maddox (1999) applied ANN algorithm to classify spectra of galaxies, and Schierscher & Paunzen (2011) also used ANN to classify stellar spectra of SDSS DR7. Besides, Hampton et al. (2017) utilized ANN to classify multicomponent emission lines with integral field spectroscopy from SAMI and S7. The emerging Deep Learning (DL) method inherits the idea of ANN, which was successfully used in many fields with big data, and greatly improves the efficiency and accuracy of spectral classification (Kim & Brunner 2016; Liu et al. 2016; Paoletti et al. 2017; Fabbro et al. 2018; Tao et al. 2018).

Convolutional neural networks (CNNs) are deep neural networks that use convolution in place of general matrix multiplication in at least one of their layers (Heaton & Jeff 2017). Different with ANNs, the regularization of CNNs take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. The major advantage of CNNs for spectral classification is that the convolution core in various scales can retain information of spectral features especially for low S/N spectra, and in other words the weak features could be enhanced by convolution. As we know that feature extraction steps for a neural network input layer will result in loss of information inevitably (Acquarelli et al. 2017).

Another issue that affects the classification performance of machine learning algorithms, including CNNs, is the imbalanced training set problem (Du et al. 2016; Zhan et al. 2018; Zhu et al. 2018). For example, the total number of O-type spectra with high S/N in LAMOST DR1 $\sim$ DR6 is only 156, which is relatively small compared with other types of stars attributing to their lifetime (Li et al. 2018; Martins 2018). The internal reaction rates of massive and hot O type stars make the gravitational field unable to be restrained (Puls et al. 1995). Unstable to exist for a long time makes the O type stars scarce to be observed, so they are too precious to be missed. In order to recognize rare objets like O type stars as many as possible in a survey data set requires the classifier to be well trained by a big balanced training data set, and it is obvious insufficient for any class which have only one hundred positive samples.

Over-sampling is a simple approach to balance training sets, which means to repeatedly sample the minority classes with relatively low proportion in the training set, however over-sampling can not better avoid overfitting of training. Another way in dealing with imbalanced training set is using simulated data (Im et al. 2016; Antoniou et al. 2017; Bowles et al. 2018; Luo et al. 2018). To increase the proportion of O type stars in the training set for CNN classifier in this paper, we propose an objective spectra producing method based on one-dimensional Generative Adversarial Network (GAN), namely Spectral GAN(SGAN), which is modified from popular two-dimensional GAN. By feeding 156 O type stars to train the SGAN, a large quantity O-type spectra can be produced to balance the training set for improving the performance of the CNN classifier.

This paper is organized as follows. Section 2 briefly introduces the spectral data used in the paper along with the spectrum preprocessing. Section 3 presents details of the methods we used including 1D CNN for spectral classification, and 1D SGAN spectra generator. Section 4 discusses the model training for CNN classifiers and the construction of balanced training data set, as well as the performance evaluation on real spectral data. Section 5 describes the application of the trained model to "UNKNOW" data set in LAMOST DR6 V1, including the identification of stellar spectra with probability cut, verification through measuring distances between classified spectra with their corresponding templates, and manual identification of O type stars. Section 6 summarizes the work of this paper, analyzes the final results, and presents challenging aspects.

2. Data and Preprocessing

The LAMOST telescope is one of the world's most efficient telescope for spectral acquisition, adopting the technologies of active optics, parallel fiber positioning and complex data analysis (Cui et al. 2012). It can obtain spectra of 4000 celestial bodies per exposure equipped with 4000 fibers, 16 spectrographs, and 32 CCD cameras. LAMOST has been running for eight years, and released more than ten million spectra in the six data release (LAMOST DR6). It is a complete astronomical data set not only because of the number of samples, but also for sky coverage, the survey volume, the sampling density, and statistical consistency (Luo et al. 2018).

2.1. "UNKNOWN" Data from LAMOST DR6

In 2017 June, the LAMOST phase one survey LAMOST-I completed. From 2017 September, the LAMOST phase two survey LAMOST-II starts which includes the new launched mid-resolution pilot survey and continuous low-resolution regular survey. The LAMOST DR6 released data from 2011 September to 2018 June including low-resolution spectra of first seven years and medium-resolution spectra for the first year. DR6 totally contains 4902 observations obtaining 11.25 million spectra, and 9.91 million of them are low-resolution spectra. Among low-resolution spectra of DR6, 9.37 million are high quality spectra with (S/N>10), and 719,651 bad quality spectra are labeled as "UNKNOWN" which were refused to classify by LAMOST pipeline. The pipeline refuses to classify a spectrum in the following two cases, either the confidence that matches the best fit template is low, or a similar chi-square matches multiple templates.

The spectra we used in this paper are internal release of DR6 (LAMOST DR6 V1), and the formal version of DR6 will be released in 2020 September. The data set of this paper is divided into four parts. For SGAN, 156 O-spectrums are

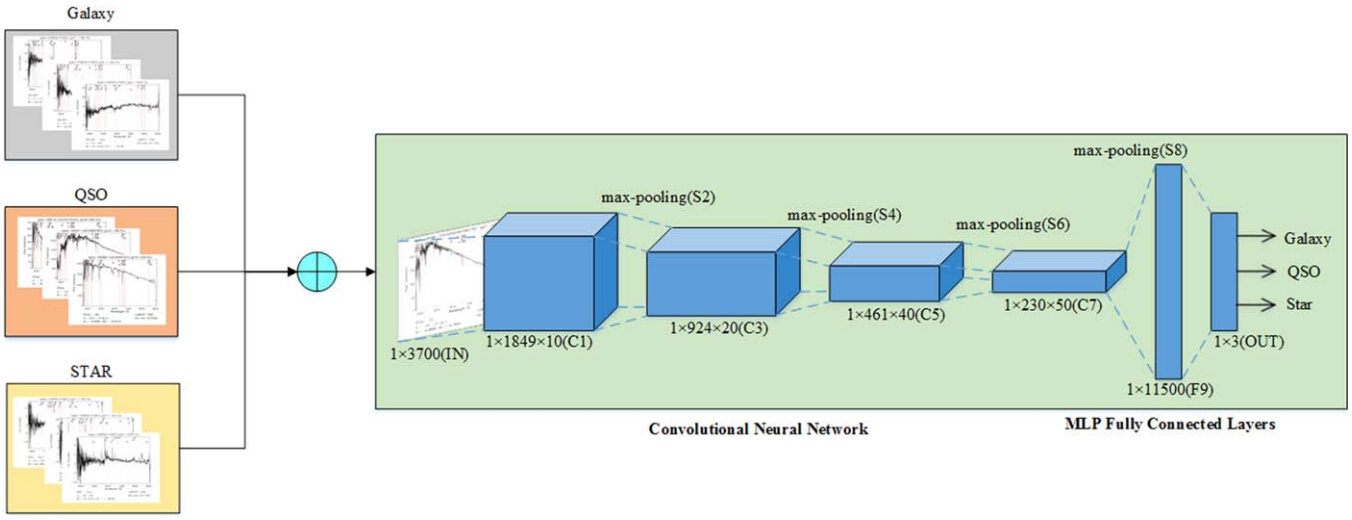

Figure 1. Architecture of the 1D CNN and training it for classifying spectra. The final MLP fully connected layers outputs a 1 × 3 vector to illustrate the type of the spectrum.
(A color version of this figure is available in the online journal.)

selected as training samples. For the first CNN, 10,500 spectra of three categories including Galaxy, QSO, and STAR are collected in the training set, and another 1500 are used as validation samples. For the second CNN, 7000 spectra of the O to M type are used as training samples (where the O-type spectrum is generated by SGAN), and there are another 2800 samples in the validation set. Finally, we collected 30,000 stellar spectra to be the test samples, among which 5000 are for each type from B to M respectively. The 5000 test samples of each type are distributed averagely in five S/N bins, i.e., 5–10, 10–20, 20–50, 50–100, and >100.

2.2. Data Preprocessing

The main preprocessing is normalization (Miyato et al. 2018). In order to ensure effective training of CNN, all the original spectra needs to be scaled to the same size using the following same normalization:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where $x_{\max}$ is the maximum value of the sample data and $x_{\min}$ is the minimum value of the sample data. It should be noticed that the normalization here only means to scale the spectra rather than continuum normalization. We have compared the classification accuracy of CNN between using continuum normalized and non normalized spectra, and there is no obvious difference, which means the spectral continuum would not affect CNN algorithm. To avoid addition error might be introduced by the step of continuum normalization, we keep the continuum in spectra.

Before scale normalization, we have to eliminate abnormal points in spectra which may lead to error in subsequent computation. The flux values of most abnormal points In LAMOST spectra have been given "0" or "NAN," and it is easy to find them and replace then by smoothing.

Then, the wavelength range of all spectra are cut in to the same window from 3700.0 to 8671.6 Å, and re-sampled to guarantee same format of the data. After the preprocess, each spectrum is an array of 1 × 3700.

3. Method

3.1. CNN Based Classifier

LeNet is a classic CNN architecture consisting mainly of convolutional layer, pooling layer, and fully connected layer Lecun et al. (1998), Ketkar (2017). The one-dimensional (1D) CNN proposed as a spectral classifier in this paper is similar to LeNet. In order to accommodate the particularity of 1D spectral data, we used different convolution kernel sizes, number of feature maps, and number of convolution layers. The network structure diagram of 1D CNN is shown in Figure 1, and the parameters of each layer are shown in Table 1.

The input layer size is 1 × 3700. A convolution kernel of a specific size performs the convolution operation to generate m feature maps,

$$h^{(m)} = \sigma\left(\sum_{i=0}^{4}\sum_{j=0}^{4} x_{i,j} \cdot \omega_{i,j}^{(m)} + b^{(m)}\right), \tag{2}$$

where $\sigma(\bullet)$ is the Relu activation function, $\omega^{(m)}$ is the weight matrix and $b^{(m)}$ is the bias vector. The formula for the Relu

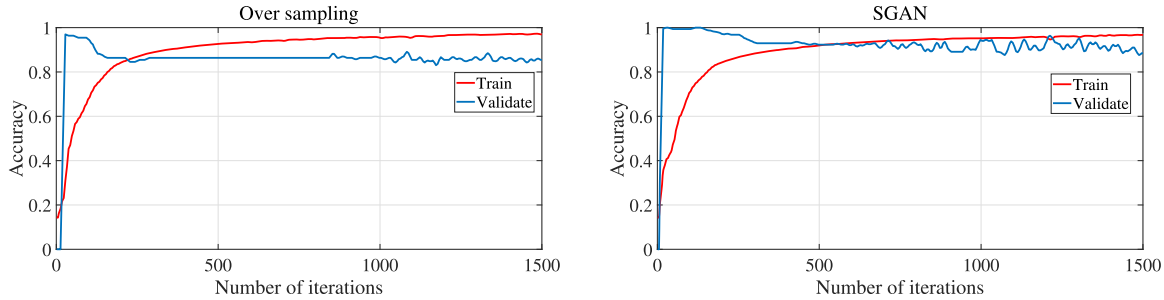

Figure 2. Comparison of classification accuracy by using different sample balance approach. The left panel shows the result of over-sampling method, while the right panel gives the accuracy of classification by using SGAN generated spectra.
(A color version of this figure is available in the online journal.)

Table 1
Parameters of Each Layer in CNN

Layer	Type	Maps	Kernel Size	Stride	Padding	Size	Activation
OUT	Fully Connected	⋯	⋯	⋯	⋯	1×3	Softmax
F9	Fully Connected	⋯	⋯	⋯	⋯	1×11500	Relu
S8	Max Pooling	50	2×2	1	VALID	1×230	⋯
C7	Convolution	50	1×2	1	VALID	1×460	Relu
S6	Max Pooling	40	2×2	2	VALID	1×461	⋯
C5	Convolution	40	1×3	1	VALID	1×922	Relu
S4	Max Pooling	20	2×2	1	VALID	1×924	⋯
C3	Convolution	20	1×2	1	VALID	1×1848	Relu
S2	Max Pooling	10	2×2	1	VALID	1×1849	⋯
C1	Convolution	10	1×3	1	VALID	1×3698	Relu
IN	INPUT	1	⋯	⋯	⋯	1×3700	⋯

function is

$$O_{i,j} = \max(0, x). \tag{3}$$

The reason for selecting the Relu activation function is mainly because of its good nonlinearity and unsaturation, which can effectively reduce the computational complexity of the network and better the converge the network. In the convolution layer, the feature map would be sampled by the sliding of the window. Using the max-pooling layer can achieve the feature compression and reduce the computational complexity. The formula is

$$O_{i,j} = \max\{h_{q,r}^{(m)}\}, \tag{4}$$

where $q, r \in (2i, 2j), (2i+1, 2j), (2i, 2j+1), (2i+1, 2j+1)$. After four convolutional and four max-pooling layers, all features are reshaped as a one-dimensional vector f and transmitted to the MLP fully connected network, i.e.,

$$y_c = \phi\left(\sum_{j=0}^{f-1} f_j \cdot \omega_{j,c} + b_c\right), \tag{5}$$

where b_c is the offset and ω is the weight matrix. Then the probability of which class the spectrum belongs to are calculated by using the SoftMax function, i.e.,

$$\phi(h_c) = \frac{e^{h_c}}{\sum_{j=1}^{P} e^{h_j}}, \tag{6}$$

The output is a one-dimensional vector representing corresponding categories of spectra, such as Galaxy, QSO and Star.

3.2. SGAN—Spectrum Generator

To balance the training data set for CNN classifier, the minority classes can be repeatedly sampled in the total training set, which is also called over-sampling. Although it is a simple way to enlarge the member of minority classes, the resulting problem of CNN over-fitting can not be well solved. In this paper, we calculate the amount of information contained in the data set by calculating the information entropy of the data set. When the amount of information is small, it is easy for the CNN to cause over-fitting, resulting in a decrease in classification accuracy. The formula for calculating information

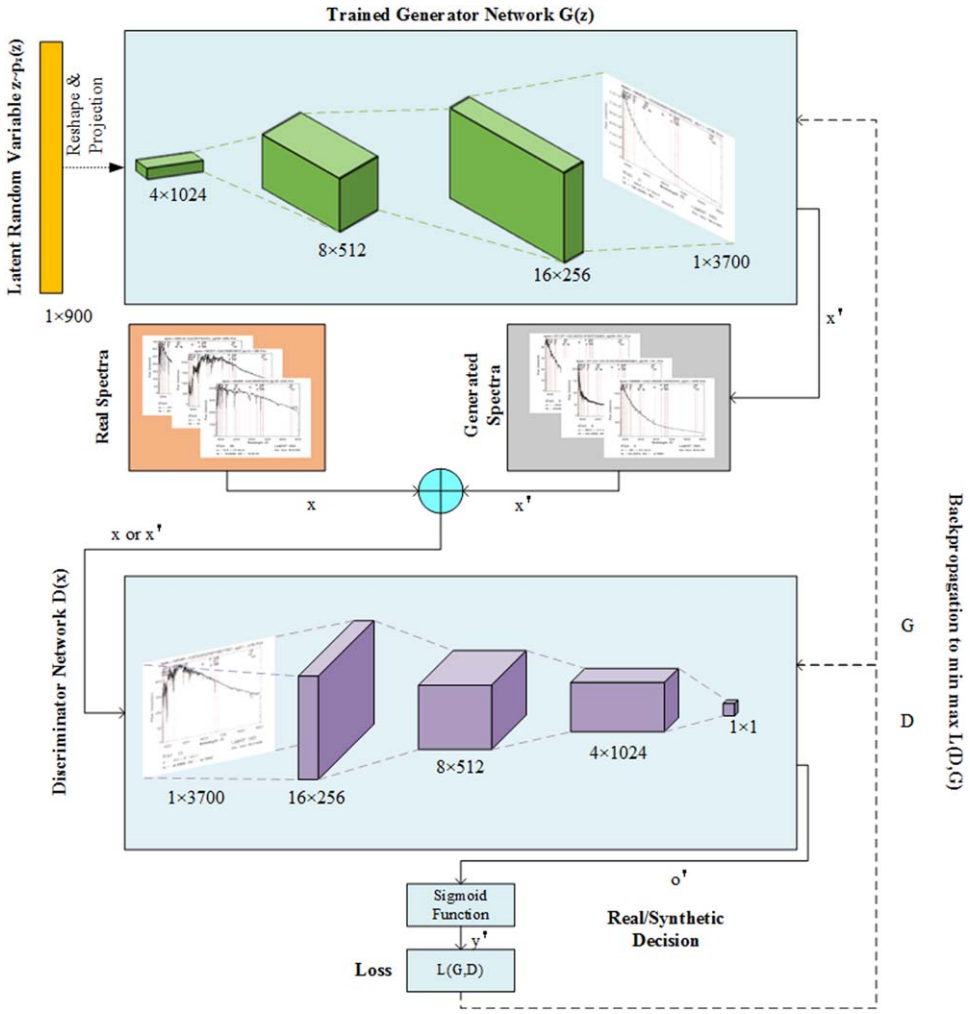

Figure 3. Architecture of the 1D SGAN. The generator uses random samples from the latent space as input to mimic real spectra in the training set as its output. The purpose of the discriminator is to distinguish the output of the generated network from real samples as much as possible. A game of both is used to generate artificial spectra.
(A color version of this figure is available in the online journal.)

entropy is as follows

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i). \quad (7)$$

Where $p(x_i)$ represents the probability that the random event X is x_i. The information entropy of the data set through the over-sampling is 5.64, while the information entropy of the SGAN generated data set is 12.29. In the information entropy calculation, the over-sampled data set is a set of 1000 sizes obtained by copying 156 original O-type spectra, and the SGAN data set consists of 1000 O-type spectra generated by SGAN.

Furthermore, it is better using SGAN generated spectra to train CNN than using over-sampled data, and the comparison is shown in Figure 2. The figure shows training accuracy and validating accuracy by using over-sampling or SGAN data to train CNN respectively. From the figure, the validation accuracy using SGAN training data is more closer to the training accuracy when the iteration converges, which means the over-fitting phenomenon in CNN training process can be effectively controlled by SGAN produced data which leads to higher classification accuracy for O-type spectra.

The network structure of the one-dimensional SGAN is shown in Figure 3. The SGAN consists of two neural networks, one is the Generator (G) and the other is the Discriminator (D), which competes with each other over the available training data to improve the spectral quality of the generated (Goodfellow et al. 2014; Radford et al. 2015).

The trained G models the underlying probability distribution p_g of the training spectra. An artificial mapping $G(z, \theta_g)$ is

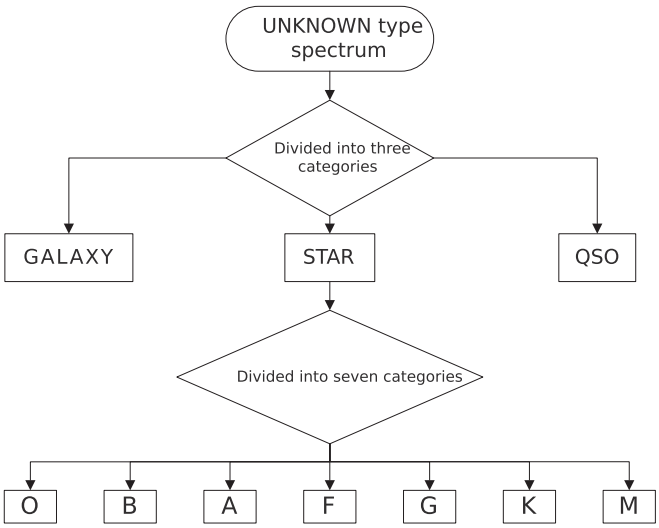

Figure 4. The low chart of the classification solution for "UNKNOWN" spectra in LAMOST DR6. First, the "UNKNOWN" spectra are divided into three types: Galaxy-QSO-Star, and then stellar spectra are divided into seven types: O, B, A, F, G, K, M.

proposed so that the input noise variable $p_z(z)$ can be mapped to real spectral data. As shown in Figure 3, a 900-dimensional noise vector z is projected to a spatially expanded convolutional representation with 1024 feature maps, then through three convolution layers the projected and reshaped noise vector z is converted to spectra of 1×3700 sampling data points.

Spectra generated by the generator G and real spectra are marked as false and true, and then input those into the discriminator network $D(x, \theta_d)$ to train the discriminator. The parameters in the generator network are not updated at this time. The discriminator then outputs an evaluation value $\hat{o}$ for each input spectrum and calculates $\hat{y}$ by formula (8)

$$\hat{y} = \frac{1}{1 + e^{\hat{o}}} \qquad s.t. \hat{y} \in (0, 1) \tag{8}$$

where $\hat{y}$ states that the spectrum is generated or real. And by the discriminating result $\hat{y}$, the parameters in the generator are updated in the next iteration process, thereby improving the capability of the generator. The competition between G and D can be expressed as the discriminator that makes the generator's maximum loss value $L(D, G)$ smaller. Formula is

$$\min_G \max_D \mathcal{L}(D, G) = \mathop{E}_{x \sim p_{\text{data}}}(x)[\log D(x)] + \mathop{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{9}$$

In each iteration process, the discriminator network updates the parameters through spectra generated by the generator, and the generator network updates the parameters through the discriminating result of the discriminator, and the two perform adversarial learning, so that the generated spectral quality is higher (Lin et al. 2017). The training set and SGAN code are available at https://github.com/Top-secreter/SGAN.

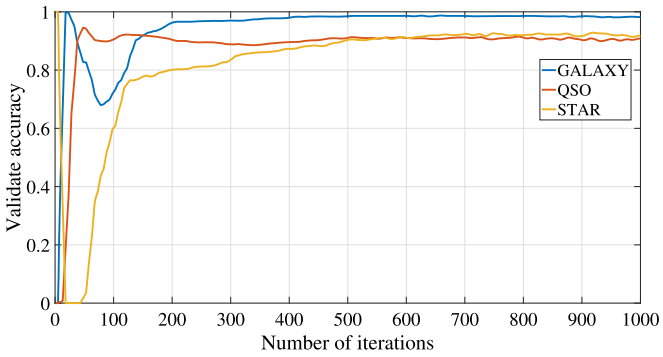

Figure 5. The accuracy curves of training process for CNN classifier which is used to divide "UNKNOWN" data into 3 types. The horizontal axis is the number of iterations and the vertical axis is the accuracy of validation set.

4. Training and Testing the Models

The whole flow chart of the classification solution for "UNKNOWN" spectra in LAMOST DR6 is shown in Figure 4. From the figure, there are two CNN classifiers, one for dividing "UNKOWN" spectra into 3 types, and another for dividing stellar candidates into 7 subclasses.

4.1. Training the CNN Model for Separating Galaxy, QSO and Star

The CNN model is trained and validated by a total of 10,500 spectra of galaxies, QSOs and stars randomly selected in LAMOST DR6, which include 3000 spectra for each type as the training set and 500 spectra for each type as the validation set. The minimum batch size for training, the number of training iterations, and the learning rate are set to 500, 1000, and 1×10^{-4} respectively. The classification accuracy curves in the training process are shown in Figure 5, and the final classification accuracy are stable to 98.2%, 90.8%, and 91.8% for galaxies, QSOs and stars respectively when the CNN converges.

4.2. Combining SGAN and CNN to Classify Stellar Spectra

We experiment different dimensional Gaussian random noise input to the SGAN, and find that the quality of generated spectra gets better with the dimension of the input noise increases. We fix the best dimension to 1×900 from experience the since more dimension would prone to mode collapse.

The noise is subjected to a fully connected three-layer strided convolution operation in the generator, and finally a batch of 1×3700 size data is obtained. The generated spectra are

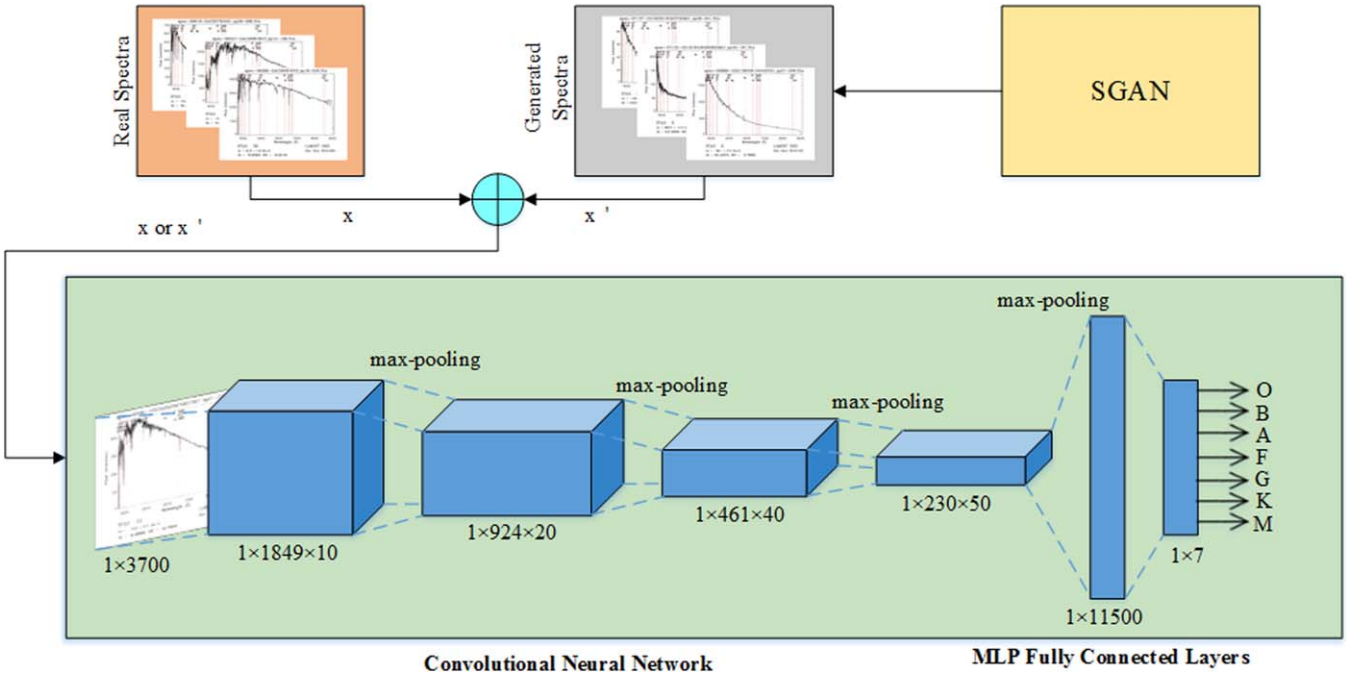

Figure 6. Architecture of the combination of CNN and SGAN. The O-type spectra generated by SGAN is mixed with the real spectra as the balanced training data set to train CNN for stellar spectral classification.
(A color version of this figure is available in the online journal.)

marked with 0 and the preprocessed O-type star spectra are marked with 1, and fed into the discriminator to train its discriminating ability. In the process of marking, we used the operation of bilateral label smoothing, which makes network convergence easier.

The discriminator's discriminating result is passed to the generator through the loss function to update the network parameters in the generator. In each iteration, we train the discriminator twice and train the generator once to ensure the stability and fast convergence of the network. We determine whether the network has reached convergence by observing the value of the loss function, as shown in Figure 7.

Since the number of O type star is much fewer than other types in DR6, the SGAN is mainly used to generate O-type samples for the CNN training set. The architecture and algorithm has been discussed in Section 3.2. After a certain number of iterative trainings, SGAN can generate O-type spectra stably. Both generated and real spectra are used to train CNN to classy stellar spectra into spectral classes of O, B, A, F, G, K, and M. The network structure of the combination of CNN and SGAN is shown in Figure 6.

We select three spectra from the official release data of LAMOST as the template spectrum of each class, and calculate the similarity between generated spectra and template spectra by calculating the Euclidean distance. The Euclidean distance between generated spectra and template spectra is taken as the

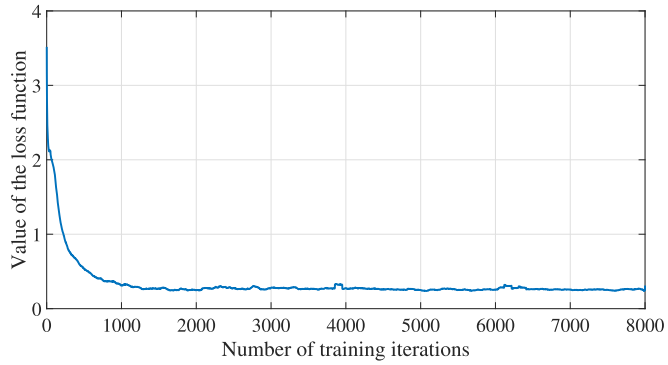

Figure 7. Loss graph of SGAN. The horizontal axis is the number of iterations, and the vertical axis is the average of the Euclidean distance of generated spectra from real spectra.
(A color version of this figure is available in the online journal.)

loss function for the generator of the 1D SGAN. As shown in Figure 7, the distance decreases with the number of iterations increases, which means that the capability of the generator is constantly increasing with the iterations increase, and generated spectra are becoming closer to real spectra constantly. When the number of iterations reaches 1500 or more, SGAN reaches convergence. Figure 8 shows an example of comparison between a generated spectra with a real spectra, and the two spectra are similar both in continuum and absorption lines,

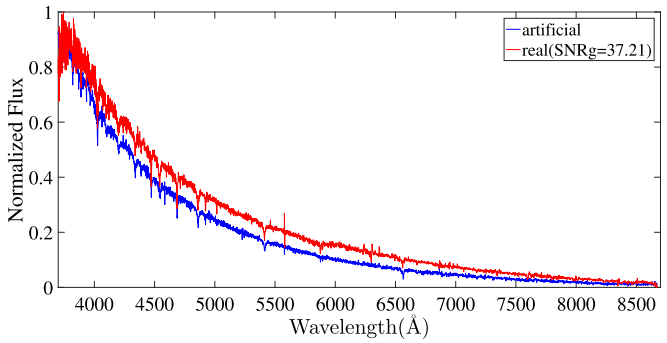

Figure 8. Comparison of a real and an artificial O-type spectrum. The horizontal axis is the wavelength, and the vertical axis is the normalized flux values. The red curve is a typical O-type real spectrum, and the blue one is generated by SGAN.

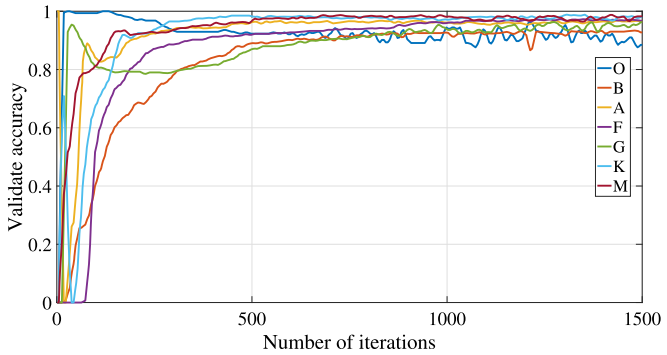

Figure 9. The accuracy curves of dividing stellar candidates into 7 types. The horizontal axis is the number of iterations and the vertical axis is the validation set accuracy.

which proves that the generated spectra can be used as training samples for CNN model.

4.3. Training and Testing the CNN Model for Stellar Classification

The training and validation sets for the CNN model are randomly selected from LAMOST DR6. For B, A, F, G, K, and M types, 1000 spectra for each are taken as training samples, and 400 spectra for each as the validation samples. For O type, 156 real spectra from DR6 and 1244 generated spectra are mixed to randomly make up 1000 training samples and 400 validation samples. The minimum batch size for training is 500, the number of training iterations is 1500, and the learning rate is 1×10^{-4}. After training completed, the classification accuracy varies with the number of iterations as shown in Figure 9, and the final classification accuracy of each type is shown in Table 2. We can see that the classification accuracy of the stellar spectra increases with the number of iterations. We choose a network with the iteration number of 1500 to save. The average classification accuracy of the stellar spectra at this time reaches 95.3%. In order to better reflect the purity and

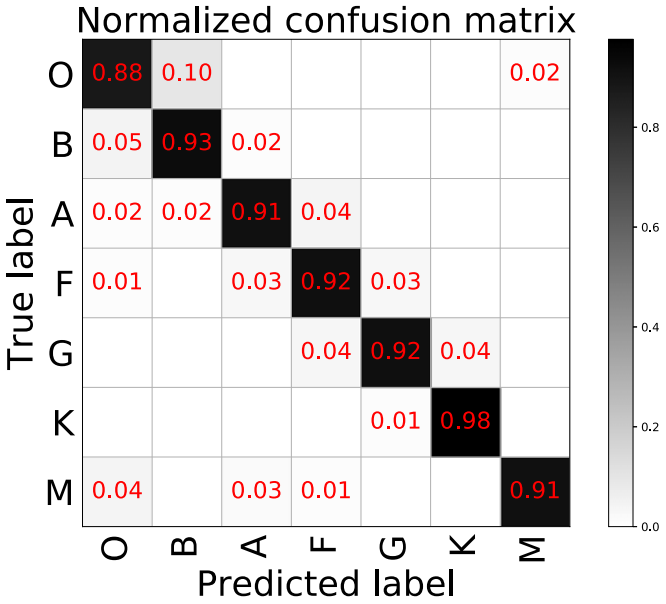

Figure 10. CNN classifies the confusion matrix of stellar spectra. (A color version of this figure is available in the online journal.)

pollution of each category, the confusion matrix of the CNN classification stellar spectra is shown in Figure 10. From the figure we can see that the spectral classification results are better concentrated on the diagonal, and the CNN has converged. We save the model at this time for the classification of the stellar spectra.

We also discuss learning parameters and classification accuracy by analyzing how the learning parameters affect the performance of the CNN. Figure 11 shows the correlation between the number of convolution layers, learning rate, kernel size, the number of feature maps, and classification accuracy. The left upper panel illustrates that the training time increases rapidly with the number of convolution layers when there are more than 4 layers in CNN. However, the classification accuracy does not increase with increasing number of network layers. So we set the number of the layers to 4. The choice of learning rate has a great influence on the algorithm convergence. In the right upper panel we can see that when the value of the learning rate is greater than 0.003, the classification performance of the CNN decreases obviously. When the learning rate is 0.0001, the network can achieve a high accuracy rate without too long training time. In the left lower panel, the network does not converge when the size of the convolution kernel of the network is greater than 64. So, we suggest that the size of the convolution kernel should be less than 32. In the right lower panel we can see that when the number of the feature maps is greater than 64, it does not improve the performance of network classification. However, as the number of feature maps increases, the training time of

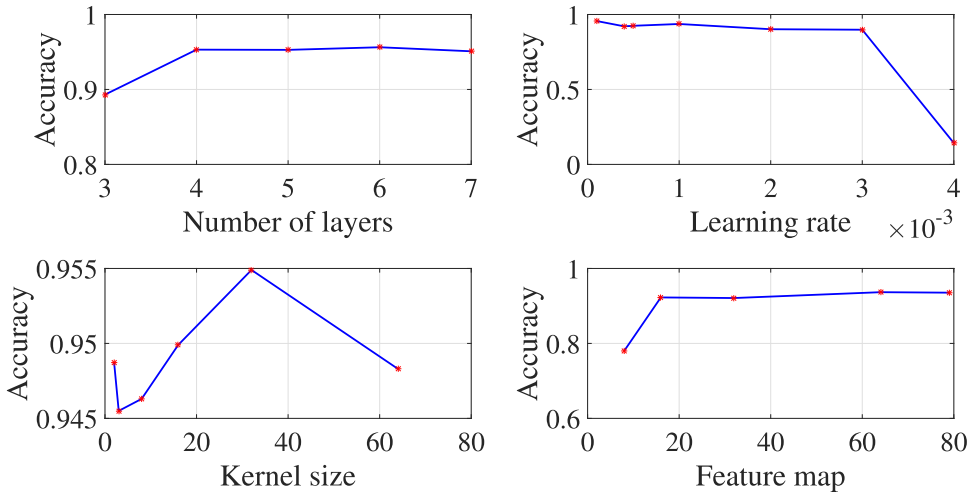

Figure 11. The correlation between the number of convolution layers, learning rate, kernel size, the number of feature maps and classification accuracy. The left panel up shows the relationship between the number of convolution layers and the classification accuracy. The right panel up shows the relationship between learning rate and the classification accuracy. The left panel down shows the relationship between kernel size and the classification accuracy. The right panel down shows the relationship between the number of feature maps and the classification accuracy.
(A color version of this figure is available in the online journal.)

Table 2
The Correct Rate of Classification of Validation Sets of 7 Types Stellar Spectra After 1500 Iterations

Star Type	O	B	A	F	G	K	M	Average
Accuracy rate	88.5%	93.3%	97.3%	95.8%	96.0%	97.5%	98.8%	95.3%

Table 3
The Correct Rate of Classification of Test Sets of 7 Types Stellar Spectra After add Gaussian White Noise

Star Type	O	B	A	F	G	K	M	Average
Accuracy rate	92.9%	71.4%	94.3%	89.3%	87.2%	95.2%	97.6%	89.7%

CNN increases significantly. So we set the number of the feature maps to 64. After analyzing the accuracy, training time, and convergence rate, the configurations are set and described in detail in Table 1.

We make a test set based on the value of the spectral S/N for testing the classification accuracy of spectra with different S/N. The B to M spectra are in five groups according to the S/N., including 5–10, 10–20, 20–50, 50–100, and greater than 100. The classification accuracy for different S/N is shown in Figure 12. We can see that when the spectral S/N is greater than 10, CNN can achieve classification accuracy better than 85%, which is a very good solution for 'UNKNOWN" spectra with S/N around 10.

In addition, we add Gaussian white noise to the observed spectra to verify the anti-interference ability of CNN. The observed spectra we used and the spectra after adding noise are shown in Figure 13. Noise is added to all the test spectra, and the classification accuracy rate is shown in the Table 3. We notice that the classification accuracy of the B-type spectra with added noise are significantly reduced, waning us that most spectral lines in the B-type star are weak and susceptible to noise. On the contrary, the accuracy of O-type spectral classification has increased, mainly because of the strong helium lines that are less affected by noise.

5. Application to "UNKNOWN" Data

5.1. Automated Identification of Stellar Spectra

Then the trained CNN classifiers are applied to recognize some "UNKNOWN" spectra as 26,761 galaxies, 7760 QSOs

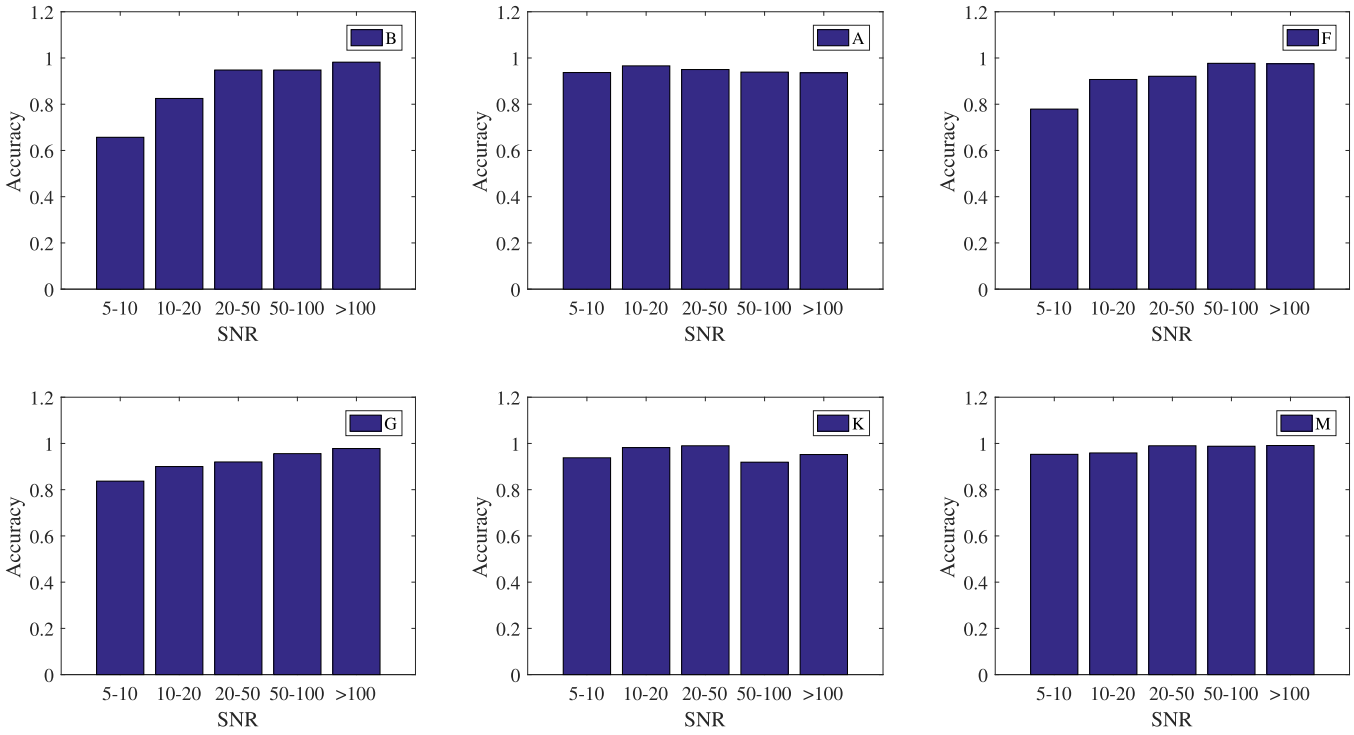

Figure 12. The accuracy of the different S/N spectra is classified by CNN. The horizontal axis is the division of the S/N. The vertical axis is the classification accuracy.
(A color version of this figure is available in the online journal.)

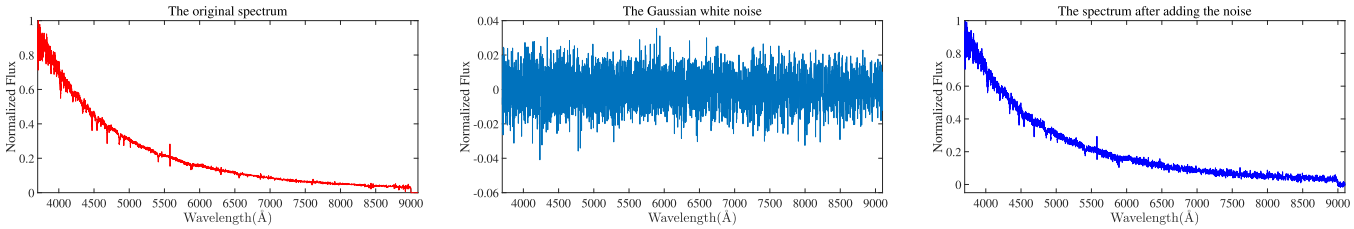

Figure 13. The left panel shows the original spectrum, the middle panel shows the added Gaussian white noise, the right panel shows the spectrum after adding the noise.
(A color version of this figure is available in the online journal.)

and 683,127 stars, and stars are further classified to 7 subclasses of O, B, A, F, G, K, and M. In order to obtain reliable results, we only retained the recognized spectra with probability larger than 99%, and 112,605 stars are finally identified. The number of identified spectra of each type is shown in Table 4, and Figures 14 shows examples of identified spectra. From the figures, we can see that the stellar spectra (blue) we find from "UNKNOWN" are consistent with the continuum of the template spectra (red), and the absorption lines can also be matched. We show the distribution of the confidence of all the "UNKNOWN" spectra in Figure 15. From the figure, the confidence of more than 66% of the spectra are above 0.5, and the spectral confidence of more than 30.8% is above 0.9. This proves the validity of the classification method.

5.2. The Limitation of the Method

In the "UNKNOWN" spectral data, there are still some spectra that we cannot distinguish, and we select three representative spectra shown in the Figure 16. The S/N of these three spectra is too low, and the value of magnitudes are too large, indicating that these stars are too dark for the telescope to obtain its valid information, which lead CNN not

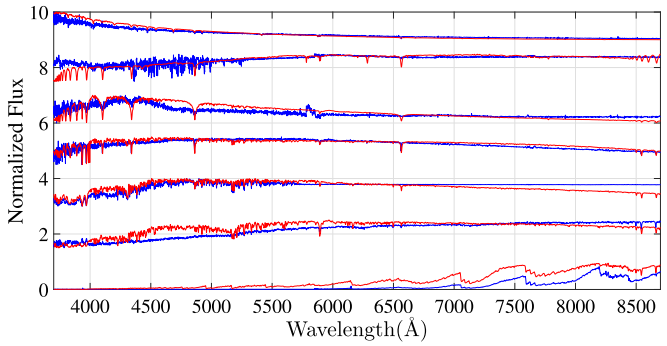

Figure 14. The examples of identified spectra of each class from O to M. The blue spectra are "UNKNOWN" spectra, and the red spectra are corresponding templates spectra. The flux in the plot is normalized.

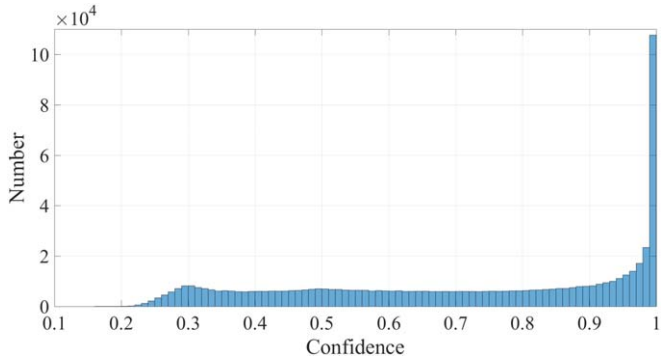

Figure 15. This is the histogram of the confidence distribution of the "UNKNOWN" spectra after CNN classification. The horizontal axis is the distribution of confidence. There are 100 confidence intervals from 0 to 1. The vertical axis is the number of spectra in each confidence interval.
(A color version of this figure is available in the online journal.)

Table 4
Number of 7 Types Spectra Classified by the Saved 1D CNN Model

Star Type	O	B	A	F	G	K	M
Number	294	3224	297	7898	661	66687	33533

unable to classify them. Of cause, there may be some other reasons to make spectra worse, such as partial loss of a certain band and instrument response. Even human can hardly judge the type of spectra.

5.3. Verification Through Average Spectra

To verify the classification result for "Unknown" data, we would compare the classified spectra with their corresponding template described in Section 4.2. To simplify the comparison, we co-add all the classified stellar spectra for each class and calculate the average spectra for the class. Using this co-added average spectra, we can augment the commonality of the spectra to compare with templates. As shown in Figure 17, we note that the continuum of the template spectra of the seven types of stellar spectra are almost identical to those of the average spectra, and the absorption lines can be well matched, which can be proved types of the classified spectra are correct.

5.4. Manual Identification

In the training set of stellar spectra, we used SGAN to generate 1,244 O-type spectra to balance the data set. Compared with over-sampling, this algorithm greatly improves the classification accuracy of O-type spectra, and we find 294 O-type spectra from "UNKNOWN" spectra by the CNN. In order to check the results of the CNN, we have carried out an artificial identification of these spectra.

We identify the O-type with He lines and Balmer lines at the wavelengths of 4009, 4026, 4101, 4121, 4144, 4200, 4340, 4387, 4471, 4541, 4686, 4713, 4861 and 6563 Å, which are shown in Figure 18. Finally, 200 O-type spectra are identified manually, and other 94 spectra are refused because of strict identification. Most of the 94 spectra have low S/Ns, and it is difficult to recognize them manually.

6. Summary

In this paper, we present adapted 1D CNN and SGAN model to classify low S/N "UNKNOWN" spectra in LAMOST DR 6. By comparing with the over-sampling method, we prove that SGAN can better augment the data set and improve the classification performance of CNN. We demonstrate that the artificial data generated by 1D SGAN can augment the real data set, thus providing a larger amount of data for the training of the DL network, and can balance the data set, thereby greatly improving the classification accuracy for minority classes. Trained 1D CNN by combining real spectra with generated spectra and obtained the best results. Through two CNN based classifiers, we classified the "UNKNOWN" spectra of the LAMOST DR6 V1 into different spectral types and subtypes, and the classification results are shown in Section 5.1. The accuracy of the classification results are verified by comparing with a traditional distance metrics. Further, we confirm O-type spectra by visually identifying the absorption lines at specific wavelengths. The classification result as a part of this work is available at http://paperdata.china-vo.org/Classification_SGAN/result.zip. The result extends the existing O-type spectral data set, and reduces the number of spectra of the "UNKNOWN" type of DR6, which confirms the feasibility of the method.

The results in this paper show that the use of simulated spectra generated by 1D SGAN can improve the distribution of training data sets, thus improving the generalization performance of the "UNKNOWN" data classification. Our work not only reduce the "UNKNOWN" spectra in LAMOST, but also extract about 200 new O-type spectra, which is a sub-catalog for very rare objects astronomers. The successful application of this semi-supervised learning algorithm in the "UNKNOWN"

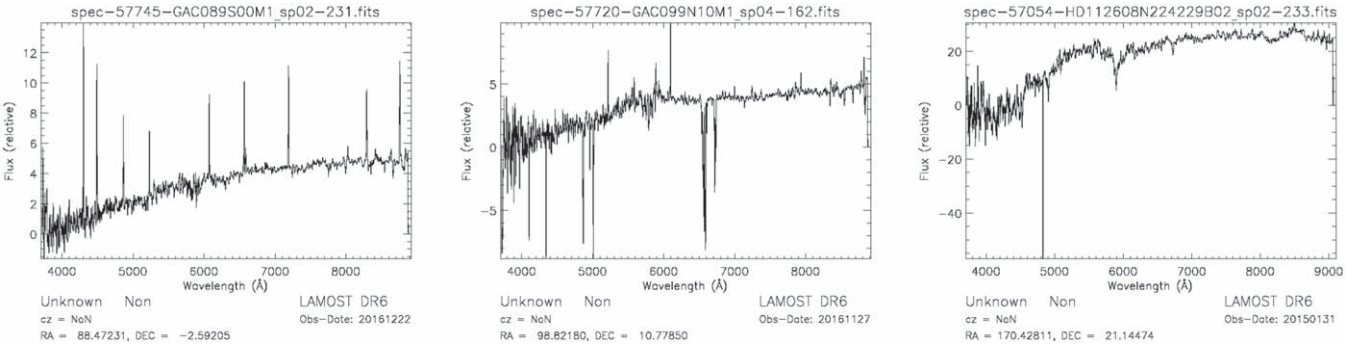

Figure 16. Failed classified spectra by the CNN. The confidence of these three spectra is 0.1997, 0.1959, and 0.1981 from left to right. The S/N is 1.82, 1.20, and 1.86 from left to right. The value of magnitude is 61.50, 63.84, 63.72 from left to right.

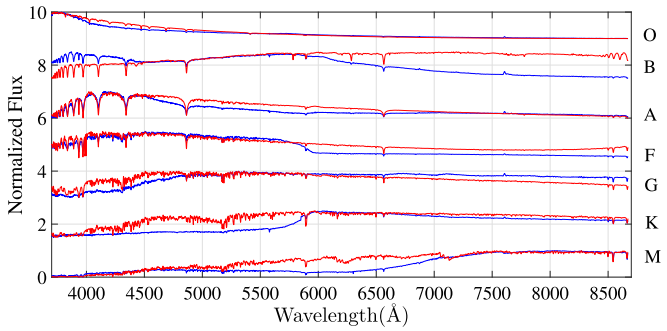

Figure 17. The blue spectra are average spectra, and the red spectra are corresponding templates spectra. The flux in the plot is normalized. From top to bottom, it is O-M.

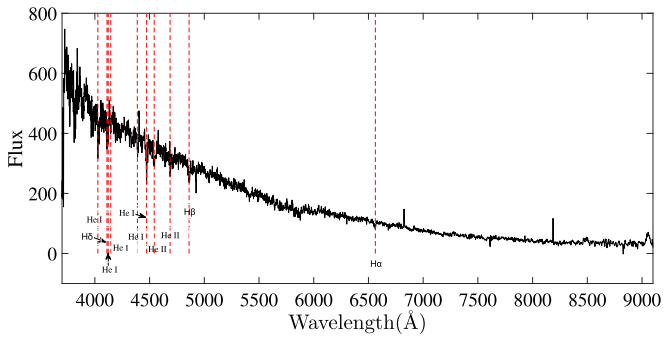

Figure 18. Auxiliary map drawn during manual authentication. The horizontal axis is the wavelength, and the vertical axis are the flux values. The red line is to observe the absorption line at this wavelength. The S/N of this spectrum is 7.68.

classification, we believe that it can get more applications in the classification of unbalanced data, especially for low S/N data.

This study is supported by the Joint Research Fund in Astronomy (Nos. U1931134, U1931209) under cooperative agreement between the National Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS), National Key R & D Program of China (No. 2019YFA0405502) and China Scholarship Council.

The Guo Shou Jing Telescope (the Large Sky AreaMulti-Object Fiber Spectroscopic Telescope, LAMOST) is a National Major Scientic Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission.

We gratefully acknowledge NVIDIA for GPU donation. Software: Numpy Oliphant (2006), Matplotlib Hunter (2007), Pandas McKinney (2011), Keras Chollet et al. (2015), Tensorflow Abadi et al. (2015).

ORCID iDs

Zi-Peng Zheng https://orcid.org/0000-0002-4519-9661
A-Li Luo https://orcid.org/0000-0001-7865-2648

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, http://tensorflow.org/
Acquarelli, J., Marchiori, E., Buydens, L. M. C., Tran, T., & Laarhoven, T. V. 2017, arXiv:1711.05512
Antoniou, A., Storkey, A., & Edwards, H. 2017, arXiv:1705.07485
Bailer-Jones, C. A. L. 1997, PASP, 109, 932
Bailer-Jones, C. A. L., Smith, K. W., Tiede, C., Sordo, R., & Vallenari, A. 2008, MNRAS, 391, 1838
Bowles, C., Chen, L., Guerrero, R., et al. 2018, arXiv:1810.10863
Carricajo, I., Manteiga, M., Rodrìguez, A., & Dafonte, C. 2004, LNEA, 1, 153
Chollet, F., et al. 2015, Keras, https://keras.io
Cui, X.-Q., Zhao, Y.-H., Chi, Y.-Q, et al. 2012, RAA, 12, 1197
Du, C.-D., Luo, A.-L., Yang, H.-F., Hou, W., & Guo, Y.-X. 2016, PASP, 128, 961
Fabbro, S., Venn, K. A., O"Briain, T., et al. 2018, MNRAS, 475, 2978
Folkes, S. R., Lahav, O., & Maddox, O. L. S. J. 1999, in Symp. Int. Astronomical Union, An ANN Approach to Classification of Galaxy Spectra for the 2DF Galaxy Redshift Survey (Cambridge: Cambridge Univ. Press), 154
Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. 2014, Adv. Nat. Inf. Proc. Syst., 3, 2672
Guo, Y. X., Luo, A.-L., Zhang, S., et al. 2019, MNRAS, 485, 2167
Hampton, E. J., Medling, A. M., Groves, B., et al. 2017, MNRAS, 470, 3395
Heaton, & Jeff 2017, Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning, Genetic Programming and Evolvable Machines, vol. 19 (Berlin: Springer)
Hunter, J. D. 2007, CSE, 9, 90

Im, D. J., Kim, C. D., Jiang, H., & Memisevic, R. 2016, arXiv:1602.05110
Ketkar, N. 2017, Introduction to Deep Learning (Berkeley CA: Apress)
Kim, E. J., & Brunner, R. J. 2016, MNRAS, arXiv:1608.04369
LaSala, J. 1994, in ASP Conf. Ser. 60, The MK Process at 50 Years: A Powerful Tool for Astrophysical Insight, ed. C. J. Corbally, R. O. Gray, & R. F. Garrison (San Francisco, CA: ASP), 312
Lecun, Y., Bottou, L., Bengio, Y., et al. 1998, Proc. IEEE, 86, 2278
Li, G.-W., Shi, J.-R, Yanny, B., et al. 2018, AJ, 863, 70
Lin, Z., Khetan, A., Fanti, G., & Oh, S. 2017, arXiv:1712.04086
Liu, Z., Song, L., & Zhao, W. 2016, MNRAS, 455, 4289
Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA, 15, 1095
Luo, H., Qi, Z.-F., & Li, J.-X. 2018, 2018 Chinese Control and Decision Conf. (CCDC), 1896–1901 (Piscataway, NJ: IEEE)
Martins, F. 2018, A&A, arXiv:1805.08267
McKinney, W. 2011, Pandas: Powerful Python Data Analysis Toolkit, http://pandas.sourceforge.net/
Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. 2018, arXiv:1802.05957
Oliphant, T. 2006, Guide to NumPy (USA: Trelgol Publishing)
Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. 2017, JPRS, 37, 1
Puls, J., Kudritzki, R. P., Herrero, A., Pauldrach, A. W. A., & Feldmeier, A. 1995, A&A, 305, 171
Radford, A., Metz, L., & Chintala, S. 2015, Comput. Sci., arXiv:1511.06434
Schierscher, F., & Paunzen, E. 2011, AN, 332, 597
Tao, Y.-H., Zhang, Y.-X., Cui, C.-Z., & Zhang, G. 2018, arXiv:1801.04839
Weaver, W. B., & Torres-Dodgen, A. V. 1995, ApJ, 446, 300
Zhan, Y., Hu, D., Wang, Y., & Yu, X. 2018, IGRSL, 15, 212
Zhu, L., Chen, Y.-S., Ghamisi, P., & Benediktsson, J. A. 2018, ITGRS, 56, 5064