Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 4, No. 3, 2022

Decomposition and Classification of Carbon Star Spectra

Yuling Zhang, Yadong Wu*

Sichuan University of Science and Engineering, Yibin, China

Abstract: Automatic classification of stellar spectra is an important research component of astronomical data processing and is the basis for studying stellar evolution and parameter measurements. As a rare kind of stellar spectra, carbon star spectra put forward more efficient and accurate requirements for classification methods. The traditional manual classification methods have the disadvantages of slow speed and low accuracy, which can no longer meet the practical needs of automatic classification of massive stellar spectra, especially low signal-to-noise ratio stellar spectra, and machine learning algorithms have been widely applied to stellar spectral classification. A distinctive feature of stellar spectra is high data dimensionality, and dimensionality reduction can not only realize feature extraction, but also reduce the computational effort, which is the first task of spectral classification. Traditional linear dimensionality reduction methods such as principal component analysis reduce the spectra only based on the variance, and different types of spectra will appear crossover after projection into the low-dimensional feature space, while streamwise learning can produce excellent classification boundaries, which will avoid overlap and facilitate subsequent classification. In view of the high dimensionality of spectral data, we investigate the distribution of spectral data in high-dimensional space and the principle of dimensionality reduction of high-dimensional linear data by stream shape learning, compare the effect of two-dimensionality reduction methods, t - SNE and principal component analysis, on spectral data, and finally analyze the experimental results and compare and validate them using various machine learning classifiers. The algorithm is implemented using Python language and Scikit - learn third-party library to perform experiments on 1000 low signal-to-noise carbon star spectra from LAMOST, and finally achieve high accuracy automatic processing and classification of the spectral data. The experimental results show that for the dimensionality reduction processing of spectral data, the t - SNE method based on stream shape learning can recover the low-dimensional stream shape structure in the high-dimensional spectral data, and after feature extraction, satisfactory classification accuracy can be achieved on the test dataset using a machine learning classifier.

Keywords: Classification of stellar spectra, Data reduction, Carbon star, Manifold learning.

1. Introduction

With the expansion of modern astronomical survey projects generating huge volumes of survey data, manual classification methods through traditional spectroscopy methods can no longer meet the needs of modern survey missions for high efficiency, high accuracy, and low labor cost. Machine learning algorithms are now widely used in astronomical spectral classification and have achieved good results. Navarro [1] used artificial neural networks to classify spectral data with different signal-to-noise ratios, and the classification results have high confidence for spectral data with low signal-to-noise ratios as well; Kheirdastan [2, 3] used probabilistic neural networks as an automatic classification tool for massive stellar spectra and obtained accurate spectral-type classification results. In addition, the classification results of stellar spectral data using an entropy learning machine are also more accurate; Chen [4] improved the efficiency of spectral classification using a restricted Boltzmann machine.

Carbon stars are rare objects, first discovered and studied by the Italian astronomer Secchi [5] in 1869. Compared with ordinary stars, carbon stars have unique physical properties, such as a higher content of carbon than oxygen in the atmosphere (C/O>1) and a spectrum characterized by strong carbon molecular bands of CH, CN, and C2, making it of great importance for the study of galactic structure, near-field cosmology, and the measurement of galactic rotation curves [6].

The carbon star spectra are classified into five types of spectroscopic spectra, C-H, C-N, C-J, C-R, and Ba, according to the Keenan [7] modified MK carbon star classification. Different types of carbon stars have different metal abundances, different galactic distribution locations, different brightnesses, different kinematic velocities, and are at different evolutionary stages, so the study of carbon star classification plays a crucial role for astronomers to study carbon stars.

LAMOST (Large Sky Area Multi-object Fiber Spectroscopic Telescope, LAMOST) [8], also known as Guo Shoujing Telescope, is China's independent innovation, the world's largest aperture of large viewport cum large aperture and spectral acquisition rate of the telescope, a 4-meter reflecting Schmidt telescope, in 20 square degrees of the focal plane with 4,000 optical fibers. In March 2020, the LAMOST team released the DR7 catalog, and in October 2021 released the DR7 v2.0 version, along with the world's largest stellar parameter catalog of about 6.91 million groups of stellar spectral parameters, which means a large number of carbon stars were discovered and the low-resolution sky area coverage.

Spectral dimensionality reduction is an important prerequisite for accurate classification. Traditional dimensionality reduction methods such as locally linear embedding [9] and linear discriminant analysis [10] have been widely applied to spectral dimensionality reduction and achieved good results. Self-encoders [11] have also been widely used for the dimensionality reduction of data.

To address the crossover problem of traditional principal component analysis in the low-dimensional space, this paper investigates the flow learning algorithm t-SNE to reduce the dimensionality of stellar spectra to produce more obvious classification boundaries, and few overlapping problems occur in the data, and the trained classifier has better results.

2. Dimensionality Reduction and Classification Methods

2.1. t-SNE

2.1.1. Sub-section Headings

t-SNE [12] is a nonlinear dimensionality reduction algorithm based on SNE, which is suitable for reducing data to 2-3 dimensions and thus facilitating visualization. In the SNE algorithm, a probability distribution among high-dimensional objects is first constructed so that similar data have a higher probability of being selected, while data with large differences have a lower probability of being selected. SNE then constructs the probability distribution of these points in a low-dimensional space so that the probability distributions between the high-dimensional and low-dimensional spaces are as similar as possible. SNE converts the high-dimensional Euclidean distance between data points into a conditional probability that represents the similarity between the data. The conditional probability $p_{j|i}$ between data sample points x_i, x_j is given by equation (1).

$$p_{j|i}=\frac{\exp\left(-\mathrm{F}x_i-x_j\ \mathrm{F}^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\mathrm{F}x_i-x_k\ \mathrm{F}^2/2\sigma_i^2\right)} \tag{1}$$

σ_i is the Gaussian variance centered on data point x_i. For the low-dimensional counterparts y_i and y_j of the high-dimensional data points x_i and x_j, a similar conditional probability can be calculated $q_{j|i}$.

$$q_{j|i}=\frac{\exp\left(-\mathrm{F}y_i-y_j\ \mathrm{F}^2\right)}{\sum_{k\neq i}\exp\left(-\mathrm{F}y_i-y_k\ \mathrm{F}^2\right)} \tag{2}$$

The goal of SNE is to minimize the difference in conditional probabilities. To compute the minimum of the conditional probability difference, SNE minimizes the KL distance by the gradient descent method. However, the cost function of SNE focuses on the local structure of the data in the mapping, and the optimization of this function is difficult to achieve, so it needs to be improved in the way of implementation.

The t-SNE uses a t-distribution in the low-dimensional space that focuses more on the long-tail distribution instead of the Gaussian distribution to represent the similarity between two points. For points with greater similarity, the distance of the t-distribution in the low-dimensional space is slightly smaller, while for points with low similarity, the distance of the t-distribution in the low-dimensional space needs to be farther. This distribution can effectively handle the outlier points in the data, i.e., anomalous data, to improve the dimensionality reduction effect.

2.2. KNN algorithm based on attribute value correlation distance

The k-nearest neighbor algorithm based on attribute-value related distance is an improved algorithm for the traditional k-nearest neighbor algorithm in terms of distance function. The algorithm first calculates the distance between the samples to be classified and the training samples of known classes using the improved distance function, and then selects the first k minimum distances, and the samples with K minimum distances are called neighbors, and determines the kind of samples to be classified according to the class confidence. The algorithm pays more attention to the statistical relevance of the data rather than just measuring the Euclidean distance between the data.

The improved distance function is the correlation distance function. The correlation coefficients of samples x_1 and x_2 are

$$\rho_{x_1,x_2}=\frac{Cov\left(x_1,x_2\right)}{\sqrt{D\left(x_1\right)}\sqrt{D\left(x_2\right)}} \tag{3}$$

Define the correlation distance between samples x_1 and x_2 as

$$d\left(x_1,x_2\right)=1-\rho_{x_1,x_2}=1-\frac{Cov\left(x_1,x_2\right)}{\sqrt{D\left(x_1\right)}\sqrt{D\left(x_2\right)}}=1-\frac{E\left(\left(x_1-Ex_2\right)\left(x_1-Ex_2\right)\right)}{\sqrt{D\left(x_1\right)}\sqrt{D\left(x_2\right)}} \tag{4}$$

The smaller the correlation distance between the samples, the greater the correlation between the two samples. The class confidence is defined as: C_r is the category, X_{test} is the sample to be classified, X_r is the number of samples belonging to C_r in the neighborhood, N is the total number of samples in the neighborhood, and N_r is the number of samples belonging to C_r in the neighborhood. N_r is the number of samples belonging to C_r in the neighboring points. $T\left(C_r,X_{test}\right)$ is the class reliability of X_{test} on C_r reliability, denoted as

$$T\left(C_r,x_{test}\right)=\frac{N-N_r}{N}\frac{1}{N}\sum_{r=1}^{N_r}d\left(x_{test},x_r\right) \tag{5}$$

The more the number of samples belonging to class C_r and the smaller the class confidence, the more likely the samples to be classified The more likely to be labeled as C_r class, the more likely to be labeled as C_r class. In summary, the algorithm based on the attribute value correlation distance The algorithm steps of the K-nearest neighbor algorithm based on the correlation distance of attribute values are as follows.

(1) Calculate the correlation distance between the training sample and the test sample.

(2) Set a suitable K value, select the K training samples with the smallest correlation distance, and calculate the number of neighboring points N_r belonging to different classes. samples, and calculate the number of neighboring points N_r belonging to different classes.

(3) Calculate the class confidence $T\left(C_r, X_{test}\right)$ of the sample to be classified and each class $T\left(C_r, X_{test}\right)$, and the class of the sample to be classified is determined as the class with the smallest class confidence.

3. Experimental Procedure and Conclusions

3.1. Data reduction

A total of 1000 carbon star spectral data obtained using LAMOST DR4 and LAMOST DR7 crossover data were used for the experimental data. After normalization, the data were downscaled to two dimensions using PCA and t-SNE algorithms.

Table 1. Carbon star type sample data

Type	Ba	C-H	C-J	C-N	C-R
LAMOST DR4	719	864	400	266	226
LAMOST DR7	669	817	297	216	212

The distribution of the data in the two-dimensional plane after PCA downscaling is shown in Figure 1.

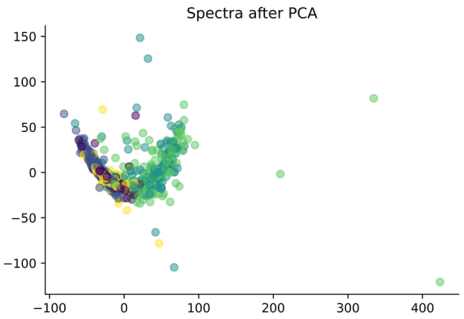

Figure 1. Distribution of data after dimension reduction by PCA

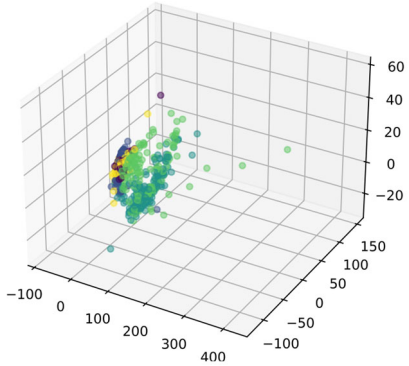

Figure 2. Distribution of data after dimension reduction by PCA

Figure 3 shows the distribution of the data after t - SNE is reduced to two dimensions. Since t - SNE is based on stream shape learning, the data after dimensionality reduction is divided into different stream shapes when In contrast, the data after dimensionality reduction by PCA are basically in the form of blocks and lines. More importantly, the Except for a very small amount of data, all data of the same type can be aggregated in one region, and There are obvious classification boundaries between different categories of data.

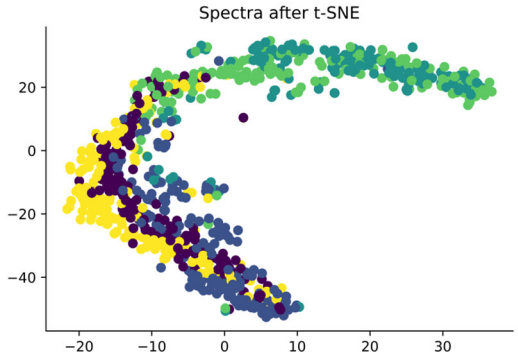

Figure 3. Distribution of data after dimension reduction by t-SNE

Comparing Figure 1 and Figure 2, it can be seen that when reducing the data to two dimensions, t - SNE is able to produce more robust classification boundaries, which provides better conditions for classification. conditions.

3.2. Analysis of results

The idea of the traditional dimensionality reduction algorithm PCA is to make the data retain the maximum variance after the dimensionality reduction. The dimensionality reduction process of PCA for spectral data is prone to data The problem of data overlap arises.

The problem is well solved in the t - SNE algorithm. According to the flow shape learning algorithm's basic idea, t - SNE can effectively extract the stream shape structure in the high-dimensional space The t SNE can effectively extract the stream structure in the high-dimensional space, thus avoiding the overlap of data in the low-dimensional space.

4. Conclusion

Carbon stars are a rare type of star that is an important part of astronomical data processing. This paper analyzes the distribution of spectral data in high-dimensional space. In this paper, we analyze the distribution of spectral data in high-dimensional space, and use the t-SNE method in stream learning to reduce the dimensionality of LAMOST carbon spectral data. The experimental results show that the t-SNE method can reduce the dimensionality of low signal-to-noise ratio stellar spectral data more effectively than the traditional PCA method. The algorithm used can significantly reduce the workload of astronomers and has a certain degree of the algorithm used can significantly reduce the workload of astronomers and has some application value.

References

[1] Navarro S G , Corradi R L M , Mampaso A . Automatic spectral classification of stellar spectra with low signal-to-noise ratio using artificial neural networks[J]. Astronomy and Astrophysics, 2012, 538:76.

[2] Kheirdastan S B M . SDSS-DR12 bulk stellar spectral classification: Artificial neural networks approach[J]. Astrophysics and space science, 2016, 361(9).

[3] Bulanov A V . Using of Ultrasound in Automated Laser Induced Breakdown Spectroscopy Complex for Operational Study of Spectral Characteristics of Seawater of Carbon Polygons[J]. Bulletin of the Russian Academy of Sciences: Physics, 2022, 86(1):S32-S36.

[4] Fuqiang C , Yan W , Yude B , et al. Spectral Classification Using Restricted Boltzmann Machine[J]. Publications of the Astronomical Society of Australia, 2014, 31:386-406.

[5] A . Schreiben des Herrn Professors Secchi an den Herausgeber[J]. Astronomische Nachrichten, 1869.

[6] T, Lloyd, Evans. Carbon stars[J]. Journal of Astrophysics & Astronomy, 2011.

[7] Keenan P C. Revised MK spectral classification of the red carbon stars[J]. Publications of the Astronomical Society of the Pacific, 1993, 105(691): 905.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces (Translation Journals style),” IEEE Transl. J. Magn. Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[8] Zhao G, Zhao Y H, Chu Y Q, et al. LAMOST spectral survey—An overview[J]. Research in Astronomy and Astrophysics, 2012, 12(7): 723.

[9] Sharma M P , Saxena R P . international journal on recent and innovation trends in computing and communication a review on non linear dimensionality reduction techniques for face recognition[J]. 2019.

[10] Park C H , Park H . A Comparison of Generalized Linear Discriminant Analysis Algorithms[J]. Pattern Recognition, 2008, 41(3):1083-1097.

[11] Mccallum A , Roweis S . Proceedings, Twenty-Fifth International Conference on Machine Learning: Preface. 2008.

[12] Gisbrecht, Mokbel, Hammer. Linear basis-function t-SNE for fast nonlinear dimensionality reduction[C]// International Joint Conference on Neural Networks. IEEE, 2012.