Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **518,** 5904–5928 (2023) https://doi.org/10.1093/mnras/stac3292
Advance Access publication 2022 November 12

Data mining techniques on astronomical spectra data – II. Classification analysis

Haifeng Yang,[1] Lichan Zhou,[1] Jianghui Cai ,[1,2]★ Chenhui Shi,[1] Yuqing Yang,[1] Xujun Zhao,[1] Juncheng Duan[1] and Xiaona Yin[1]

[1]*School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China*
[2]*School of Computer Science and Technology, North University of China, Taiyuan 030051, China*

Accepted 2022 November 7. Received 2022 November 3; in original form 2022 April 1

ABSTRACT

Classification is valuable and necessary in spectral analysis, especially for data-driven mining. Along with the rapid development of spectral surveys, a variety of classification techniques have been successfully applied to astronomical data processing. However, it is difficult to select an appropriate classification method in practical scenarios due to the different algorithmic ideas and data characteristics. Here, we present the second work in the data mining series – a review of spectral classification techniques. This work also consists of three parts: a systematic overview of current literature, experimental analyses of commonly used classification algorithms, and source codes used in this paper. First, we carefully investigate the current classification methods in astronomical literature and organize these methods into ten types based on their algorithmic ideas. For each type of algorithm, the analysis is organized from the following three perspectives. (1) their current applications and usage frequencies in spectral classification are summarized; (2) their basic ideas are introduced and preliminarily analysed; (3) the advantages and caveats of each type of algorithm are discussed. Secondly, the classification performance of different algorithms on the unified data sets is analysed. Experimental data are selected from the LAMOST survey and SDSS survey. Six groups of spectral data sets are designed from data characteristics, data qualities, and data volumes to examine the performance of these algorithms. Then the scores of nine basic algorithms are shown and discussed in the experimental analysis. Finally, nine basic algorithms source codes written in python and manuals for usage and improvement are provided.

Key words: methods: data analysis – techniques: spectroscopic – software: data analysis.

1 INTRODUCTION

Classification of astronomical spectra is an essential part of astronomical research. It can provide valuable information about the formation and evolution of the Universe. With the implementation of sky survey projects (Zhao et al. 2012; Liu, Zhao & Hou 2015a), a large number of methods have been applied to automatically handle various astronomical classification tasks (Luo, Zhang & Zhao 2004; Luo et al. 2013; Baron 2019; Yang et al. 2020, 2021, 2022c, b; Cai et al. 2022). However, classification methods achieve different results on different data, so it is difficult to evaluate the classification performance and determine the application scenarios.

In this paper, we investigate lots of classification methods on astronomical spectra data and organize them into ten types. Each type of them is displayed based on its usage frequencies in astronomical tasks. And we mainly discuss its application scenarios, main ideas, merits, and caveats. Then, we construct six collections of data sets to provide a unified measurement platform. For the astronomical classification tasks (A/F/G/K stars classification, star/galaxy/quasar classification, and rare object identification), we construct data sets from three criteria including data characteristics, signal-to-noise ratio (S/N), data volumes. Then we compare the performance of nine basic classification methods on the aforementioned data sets and give an objective appraisal of the classification results. Besides, the source codes of each testing algorithm help researchers to study further and a brief manual about usage and revision tips of our program is provided in this work.

The rest of this paper is organized as follows. In Section 2, classification methods on astronomical spectra data are briefly introduced from application scenarios, main ideas, merits, and caveats. In Section 3, experiments on three tasks of A/F/G/K stars classification, star/galaxy/quasar classification, and rare object identification are carried out. Section 4 represents python source codes of the above experiments and a manual about how to use and revise our codes. Finally, a discussion is drawn and our future work is discussed in Section 5.

2 INVESTIGATION OF CLASSIFICATION METHODS ON ASTRONOMICAL SPECTRA DATA

The commonly used classification methods on astronomical spectra are shown in Fig. 1. Each type of methods has its own characteristics and applicable data sets. And some of them have been widely used for spectral classification, like template matching, K-nearest neighbour (KNN) based classification algorithms, and support vector machine (SVM) based classification algorithms, but some of them

★ E-mail: jianghui@tyust.edu.cn

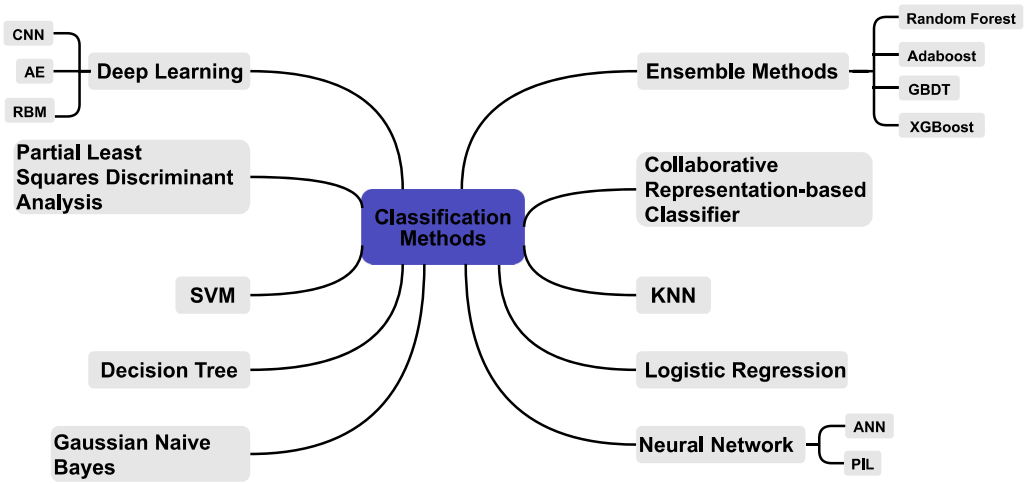

Figure 1. Classification methods on astronomical spectra data. We pay more attention on main ideas, advantages, caveats, and application scenarios of these methods.

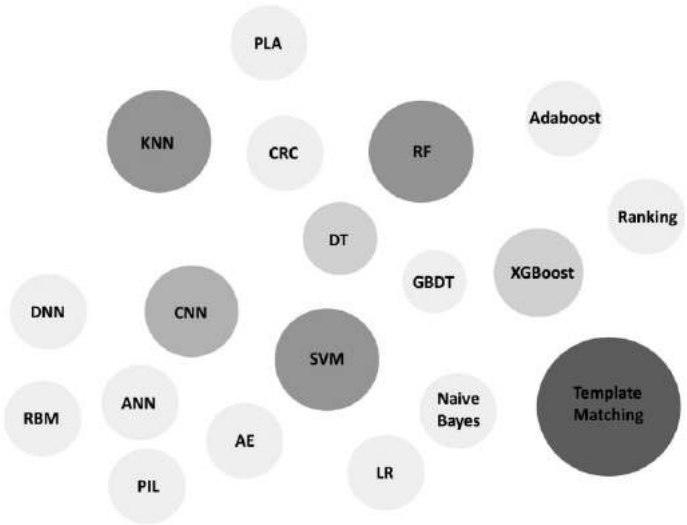

Figure 2. Classification methods on astronomical spectra data. The size of circles means the usage frequencies of each type of algorithm in our paper. And the colour of circles is consistent with their sizes. That is, bigger and deeper circles mean that this type of algorithm is more frequently applied in astronomical tasks.

are rarely used, like logistic regression (LR) based classification algorithms and collaborative representation based classifier (CRC) (Fig. 2). Here, for each type of investigated methods, we analyse its application scenarios on astronomical spectra and give some objective appraisals. Then we introduce the main ideas, advantages, and caveats of these methods.

2.1 Template matching

Template matching is a flexible and relative straightforward technique. The classification process of template matching is to build a template data base for each class, then divide the unknown data into the most similar template data (Rosenfeld & Vanderbrug 1977). In astronomy, template matching matches spectral lines with templates and there is no training stage. So it has been widely applied in celestial object classification, redshift estimation, stellar parameters estimation, and other projects (Lupton et al. 2002; SubbaRao et al. 2002; Zhao et al. 2012; Liu et al. 2015a; Westfall et al. 2019). Table 1 shows the main astronomical spectral investigations of template matching.

Template matching is often used to classify stars, galaxies, and quasars and further analyse other properties of spectra. Duan et al. (2009) used spectral line matching to identify the observed spectra class and achieved a high accuracy about 92.9 per cent, 97.9 per cent, and 98.8 per cent for stars, galaxies, and quasars, respectively. They also obtained a byproduct: high precision of redshift. Gray & Corbally (2014) used template matching for Morgan-Keenan (MK) classification and built an expert computer program imitating human classifiers. It was automatic and had comprehensible results. Wang et al. (2018) used the line intensity to classify spectra (Martins 2018; Wang 2019).

Template matching is also used to find peculiar objects like supernovas, M dwarfs, B stars, and M giants, Double-peak emission line galaxies (Zhong et al. 2015b, a; Sako et al. 2018; Maschmann et al. 2020; Ramírez-Preciado et al. 2020). Zhong et al. (2015b) applied a template-fit method to identify and classify late-type K and M dwarfs from LAMOST. 2612 late-K and M dwarfs were identified which can help researchers to investigate the chemokinematics of the local Galactic disc and halo. Maschmann et al. (2020) used two Gaussian functions to fit the emission lines to find double-peak

Table 1. Investigations of template matching on astronomical spectra data.

Merits	Caveats	References
Straightforward and simple	Poor performance on low-quality spectra	Rosenfeld & Vanderbrug (1977), Duan et al. (2009), Du et al. (2012), Ramírez-Preciado et al. (2020), Wang (2019), Almeida et al. (2010), Martins (2018), Wang et al. (2018), Li et al. (2016), Juvela (2016), Sako et al. (2018), Zhong et al. (2015a), Zhong et al. (2015b), Bolton et al. (2012), Wei et al. (2014), Gray & Corbally (2014), Masters & Capak (2011), Khorrami et al. (2021), Kesseli et al. (2017), Cotar et al. (2019), Agnello (2017), Zhang et al. (2016), Saez et al. (2015), Karpov, Malkov & Zhao (2021), Gao et al. (2019)
Applied to stellar spectra, rare objects, etc[1]	Spectra without templates cannot be classified well	
Fast because of without training stage	Poor performance on unbalanced data	

Note. [1] Subtypes of O star, Subtypes of B star, galaxy/others, etc.

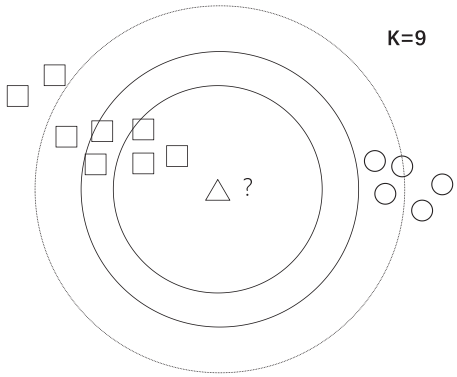

Figure 3. Process of KNN. The middle triangle is the object needed to be predicted. Rectangles and circles are two types of known objects. The dashed circle needs to be enlarged to find k neighbours. $K = 9$: three circles and six rectangles are the triangle's k neighbours.

candidates and finally they found 5663 double-peak emission line galaxies at z < 0.34. Meanwhile, there is an important issue for rare object identification using template matching, that is, classifiers require sufficient high-quality spectral templates. In order to obtain ample qualified rare templates, researchers tried to construct new templates (Wei et al. 2014; Kesseli et al. 2017).

Template matching has been widely used in lots of surveys. However, some spectra are of low quality, template matching cannot obtain precise results on redshift estimation, stellar parameters estimation, and classification (Podorvanyuk, Chilingarian & Katkov 2015). Hence, for the inferior quality spectra, other machine learning algorithms like SVM based classification algorithms and artificial neural network (ANN) based classification algorithms are employed to get robust results. The other defect of template matching is that, for rare objects, we do not have enough samples to get representative template spectra. So rare objects are often misclassified.

2.2 K-Nearest neighbour based classification algorithms

K-Nearest Neighbor (KNN) based classification algorithms assign labels to the target based on the majority labels of its K closest objects. More in depth explanations of KNN based classification algorithms can be found in Zhang & Zhou (2007), Deng et al. (2016). The main ideas are shown in Fig. 3. They are intelligible and their time complexity is linear to the data volume. Taking these into consideration, KNN based classification algorithms have been used to classify astronomical spectra and combined with other methods to improve classification accuracy. Table 2 displays the major astronomical applications of KNN based classification algorithms.

KNN can be used for stellar classification. Brice & Andonie (2019b) used KNN and random forest (RF) for MK classification of stellar spectra. Considering high dimensional spectra data, they extracted absorption lines of spectra to reduce the time complexity. The results showed that KNN had a shorter training time but a longer testing time than RF. KNN could obtain the same accuracy as RF when using hybrid methods or oversampling balancing techniques. But for O-type stars which are few in the data sets, KNN performed poorly. This is a common phenomenon in most classification applications, that is, it is hard to get good classification results in unbalanced data sets.

For complex spectral classification tasks, it is not a good choice to only use the basic KNN based classification methods. Because from the comparison results of different classification methods, researchers found that good results were often produced by SVM or RF, rather than KNN (Pérez-Ortiz et al. 2017; Arsioli & Dedin 2020; Xiao-Qing & Jin-Meng 2021). To obtain better results, some improvements to KNN were also proposed, like KNN-DD to detect known outliers (Borne & Vedachalam 2012) and ML-KNN: a lazy learning approach to multilabel learning (Zhang & Zhou 2007). In addition, many researchers combined KNN with other methods to reduce the misclassification rate, like SVM + KNN to correct some prediction errors (Peng et al. 2013). And its classification accuracy of quasars reached 97.99 per cent.

KNN based classification algorithms are arguably simple and efficient machine learning algorithms. And they have been demonstrated to be competitive methods because of the high accuracy under the premise of their simplicity and rapidness (Fushiki 2011; Guzmán et al. 2018; Sookmee et al. 2020). They use Euclidean distance to measure the similarity of data and perform better on low dimensional data. After pre-processing high dimensional spectra, KNN based classification algorithms can also be applied in astronomy, such as star/galaxy/quasar classification and classification of small radial velocity objects. However, from the investigated researches, KNN based classification algorithms mainly suffer from three disadvantages: (1) the only hyperparameter K is difficult to determine. (2) KNN based classification algorithms are ineffective for star classification because of the misclassification between adjacent classes. (3) Unbalanced data are another challenge for KNN based classification algorithms. Recently, some algorithms like Synthetic Minority Oversampling Technique (SMOTE) have been employed to adjust the data volume distributions to solve the third issue.

Table 2. Investigations of KNN based classification algorithms on astronomical spectra data.

Merits	Caveats	References
Accurate and fast on proper features[1]	Generally, RF>SVM>KNN	Brice & Andonie (2019b), Arsioli & Dedin (2020),
Combined with other methods[2] to improve accuracy	limited to large redshift objects	Peng, Zhang & Zhao (2013), Bu et al. (2019),
Applied to stellar spectra and subtypes classification[3]	Misclassification on F, G, K stars	Akras et al. (2019), Xiao-Qing & Jin-Meng (2021), Pérez-Ortiz et al. (2017), Sookmee et al. (2020)

Notes. [1] Features extracted by CNN; astronomical specific information.
[2] SVM, CNN, Decision tree, etc.
[3] MK classification, star/galaxy/quasar classification, Hot subdwarfs, symbiotic stars, Be stars, LSP/HSP, etc.

Table 3. Investigations of SVM based classification algorithms on astronomical spectra data.

Merits	Caveats	References
Accurate and fast on proper features[1]	Stellar loci is better than SVM on MK classification	Liu et al. (2015b), Guzmán et al. (2018), Arsioli & Dedin (2020), Qu et al. (2020), Liu et al. (2019), Govada, Gauri & Sahay (2015),
Optimizations of SVM to improve accuracy[2]	1D SCNN is better than SVM on stellar classification	Fuqiang et al. (2014), Barrientos, Solar & Mendoza (2020), Solarz et al. (2012), Tsalmantza et al. (2012),
Applied to unbalanced and large-scale data sets	Limited to large redshift objects	Kou, Chen & Liu (2020), Liu (2021), Małek et al. (2013), Peng et al. (2013),
Applied to stellar spectra and subtypes classification[3]	Need sufficient valid samples	Solarz et al. (2017), Liu & Zhao (2017), Yude et al. (2013), Xiao-Qing & Jin-Meng (2021), Dong & Pan (2020), Kong et al. (2018), Bu et al. (2019), Liu, Song & Zhao (2016)

Notes. [1] Multifrequency, colour space, spectral lines, etc.
[2] Within-Class Scatter and Between-Class Scatter (WBS-SVM), OCSVM, Twin Support Vector Machine (TWSVM).
[3] MK classification, LSP/HSP, K/F/G stars, Type IIP/IIL Supernovae, etc.

2.3 Support vector machine based classification algorithms

Support vector machine (SVM) based classification algorithms are binary classifiers that learn a boundary from the training data to classify two types of data. And multiple binary SVM classifiers can be integrated into a multilabel classifier. Generally, the classification precision and robustness of SVM based classification algorithms are relatively superior to other single classifiers (non-ensemble algorithms). Table 3 shows the main astronomical researches of SVM based classification algorithms.

Spectral classification is a common astronomical task for SVM based classification algorithms (Liu et al. 2015b, 2018; Guzmán et al. 2018; Tao et al. 2018; Brice & Andonie 2019a; Barrientos et al. 2020; Liu 2021). Solarz et al. (2012) used the infrared information to separate galaxies from stars and the accuracy reached 90 per cent for galaxies and 98 per cent for stars. Małek et al. (2013) trained an SVM classifier to classify stars, active galactic nucleus (AGN) and galaxies using spectroscopically confirmed sources from the VIPERS and VVDS surveys. In the stellar spectral classification, A stars and G stars can be identified easily, while it was hard to identify O, B, and K stars. Because the differences in the spectral features between late B type and early A type stars or between late G and early K type stars were very weak (Liu et al. 2015b). Dong & Pan (2020) used SVM and cascaded dimensionality reduction techniques to classify spectra, which is better than principal component analysis (PCA) or t-distributed stochastic neighbour embedding (T-SNE).

In addition to classification, SVM based classification algorithms can also be used for peculiar spectra identification (Qu et al. 2020). More depth details of rare objects such as carbon stars and variable objects can be found in Gigoyan et al. (2012), Green (2013), Baran et al. (2021), Maravelias et al. (2022), Kong et al. (2018), Kou et al. (2020), Solarz et al. (2017), Qu et al. (2020). Solarz et al. (2020) detected anomalous in the mid-infrared data using one-class SVM. Among the 36 identified anomalous, 53 per cent of them were low redshift galaxies, 33 per cent were particular quasi-stellar objects (QSOs), 3 per cent were galactic objects in dusty phases of their evolution, and 11 per cent were unknown objects. The main problem in this task is that the number of some types of rare samples is far smaller than normal samples. So the classification model cannot identify the rare classes well. There are also many approaches to solve this problem, like data augmentation, oversampling, etc. Liu & Zhao (2017) proposed an entropy based methods for unbalanced spectral classification. And the performance was better than using KNN and SVM directly.

SVM based classification algorithms are binary classifiers with rigorous mathematical theory. They try to find the optimal separating hyperplane to divide data into two categories (Fig. 4). For multilabel classification, One-VS-One (OVO), One-VS-All (OVA), and Directed Acyclic Graph (DAG) are the main tactics to train different classifiers. There are two important tricks of SVM based classification algorithms. One is soft margin which uses a robust partition boundary to separate two types of data and tolerates the

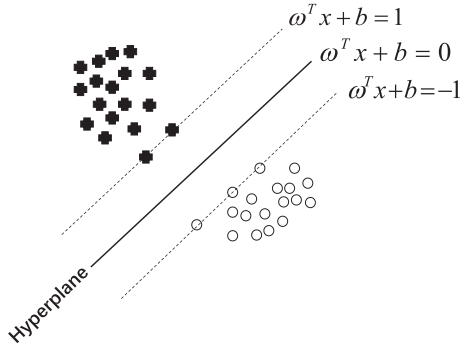

Figure 4. Process of SVM. The black circles and white circles are two types of unknown groups. The solid line is the hyperplane to separate groups. The proper gap between paralleled dashed lines and solid lines can avoid overfitting and underfitting.

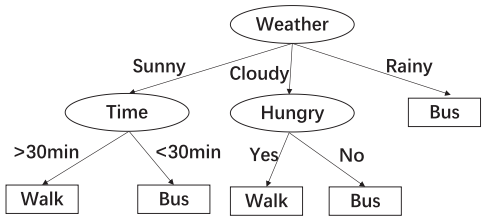

Figure 5. Main idea of decision tree. ID3, C4.5, and CART are three common decision trees, the principle of them are similar except for the different node-split tactics (Information Gain, Information Gain Ratio, and Gini Index). Ellipses represent the split nodes for classification, and rectangles represent the final multiple classification results.

misclassification of some abnormal data. The other one is kernel function which can map linearly inseparable data into a linearly separable high-dimensional space. However, SVM based classification algorithms adopt kernel matrix to measure the similarity of samples. So the computation time and space are two vital issues for classifiers on large amounts of data.

SVM based classification algorithms are promising classification methods that have a convincing theory and robust results. They have attracted a good deal of attention due to their high accuracy in multidimensional space and already have been applied to astronomical spectral classification, such as star/galaxy/quasar classification, stellar spectral classification, and novelty detection. However, the time complexity of SVM based classification algorithms is exponentially related to the training size. So, it is indispensable to pre-process astronomical spectra to reduce the training time.

2.4 Decision tree based classification algorithms

Decision tree (DT) based classification algorithms (Quinlan 1996) are essential in machine learning algorithms. Their leaves represent classification results and internal nodes of branches are regarded as criteria for distinguishing objects. The graphical representation of decision tree is shown in Fig. 5. Decision tree and its variants have been applied in astronomy and many other fields (Li 2005; Zhao & Zhang 2008; Bae 2014; Czajkowski, Grześ & Kretowski 2014). Table 4 shows the main astronomical investigations of decision tree based classification algorithms.

Decision tree based classification algorithms have been widely used for astronomical classification due to their good interpretability of classification results. Here are some examples of decision tree for astronomical classification. Morice-Atkinson et al. (2018) explored the classification boundaries of star and galaxy through decision tree. This visualized the classification process of the star–galaxy and helped astronomers understand the decision rules of celestial classification. Franco-Arcega, Flores-Flores & Gabbasov (2013) used parallel decision trees to classify different types of objects and evaluated the performance of classification results. Vasconcellos et al. (2011) applied 13 different decision tree algorithms to analyse the classification performance of star/galaxy, and the functional tree algorithm yielded the best results.

According to the astronomical researches using decision tree based classification algorithms, there are three tips to improve the classification performance. First, effectively pre-processing the raw observational spectra will assist and speed up the classification, such as noise reduction and data compression. Secondly, extracting valid features is also important. Most familiar approaches normalize and standard spectra data by prevalent methods without additional operations (Vasconcellos et al. 2011; Pichara et al. 2016), yet these simple approaches will have high computational costs and could not improve classification accuracy effectively. So other valid features may be better to improve classification performance, such as line indices and astronomical specific features. Thirdly, searching for appropriate methods is another vital approach to improving classification performance. Compared with other typical methods, RF performed best both on accuracy and time consuming in Xiao-Qing & Jin-Meng (2021), Brice & Andonie (2019b), Flores et al. (2021), etc. Alternatively, integration of decision tree and other conventional classification methods can enhance the superiority of feature selection and results interpretation, respectively (Ivanov et al. 2021). However, heterogeneous data, large redshift objects, other stellar parameters regression, and misclassification are still challenges for decision tree in astronomical research. These need to be solved in the future.

Iterative Dichotomiser 3 (ID3), C4.5, and Classification and Regression Trees (CART) are three widespread methods based on decision tree. ID3 adopts Information Gain (IG) as the node selection criterion for classification. While C4.5 chooses Information Gain Ratio (IGR) to alleviate the flaws of ID3 (IG: discrete data, incomplete attribution, overfitting, etc). Another upgraded method is CART which can be used for both classification and regression. It employs the Gini Index as a node selection standard instead of Information Entropy. The main advantage of decision tree based classification algorithms is interpretability of results, which is very helpful for astronomers to analyse the features of astronomical objects. And the disadvantage is that we often obtain a complex model which will be overfitting on the training data. So pruning parameters is always required to reduce overfitting.

2.5 Ensemble learning classification algorithms

Ensemble learning (Freund & Mason 1999) combines multiple weak classifiers into a strong classifier to solve a task together. Generally, ensemble learning methods are sorted into bagging methods decreasing variance, boosting methods reducing deviation, and stacking methods increasing prediction accuracy. Compared with the single decision tree, ensemble learning methods are more often used in astronomy.

Bagging usually trains different models with various training sets respectively and chooses one strategy to unify consequences. The principle of bagging is shown in Fig. 6. Random Forest is the most notable bagging method which consists of several unrelated decision trees. Fig. 7 describes the principle of Random Forest. RF can be

Table 4. Investigations of DT based classification algorithms and ensemble learning on astronomical spectra data.

Merits	Caveats	References
Results interpretability and predicted probability	Limited to large redshift objects	Pichara, Protopapas & León (2016), Akras et al. (2019), Flores, Corral & Fierro-Santillán (2021), Xiao-Qing & Jin-Meng (2021), Vasconcellos et al. (2011), Morice-Atkinson, Hoyle & Bacon (2018), Clarke et al. (2020),
High accuracy on stellar and subtypes[1]	Misclassification on G,F,K stars	Pattnaik et al. (2021), Li, Lin & Qiu (2019), Bai et al. (2019), Arsioli & Dedin (2020), Liu et al. (2019), Yi et al. (2014),
Extract feature well for stellar and subtypes [2]	Adjust parameters manually	Hosenie et al. (2020), Li et al. (2019), Baqui et al. (2021), Reis et al. (2018), Brice & Andonie (2019b),
Classify spectra with missing values and noise	Poor performance on unbalanced data	Pérez-Ortiz et al. (2017), Ivanov et al. (2021), Tao et al. (2018), Kyritsis et al. (2022), Guo et al. (2022), Brice & Andonie (2019a), Maravelias et al. (2022), Hou et al. (2020), Hu et al. (2021), Zhang, Zhao & Wu (2021), Yue et al. (2021), Sookmee et al. (2020)

Notes. [1] Star/galaxy/quasar classification, MK classification, LSP/HSP, M star/others, etc.
[2] MK classification, stellar subtypes, M subtypes, etc.

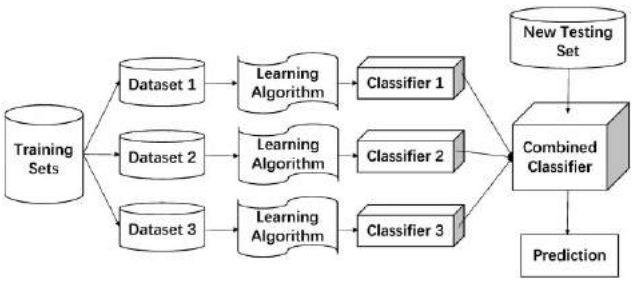

Figure 6. Main idea of bagging. The left cylinders are training sets to train different base classifiers using learning algorithms. The upper cylinder is the new testing set to test the combined classifier which is made up of a collection of base classifiers. And the strong classifier is generated by voting for classifier i ($i = 1, 2, 3$ in Fig. 6).

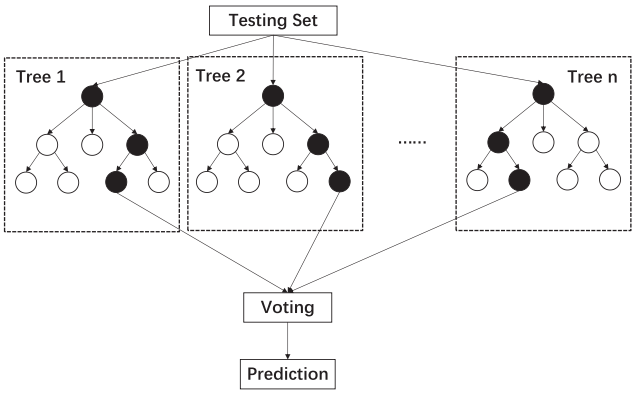

Figure 7. Main idea of random forest. Tree i ($i = 1, 2, ..., n$) are decision trees trained by certain strategies. The black circles are the basis nodes of classification. The voting methods adopt the majority or average results of tree i ($i = 1, 2, ..., n$) as the final results.

applied to classification, clustering, regression, and outlier detection due to its high accuracy and adaptability of high dimensional data sets.

RF is a robust classifier for spectral classification (Yi et al. 2014; Biau & Scornet 2016; Morice-Atkinson et al. 2018; Bai et al. 2019; Brice & Andonie 2019a, b; Liu et al. 2019; Li et al. 2019; Hosenie et al. 2020; Baqui et al. 2021). Clarke et al. (2020) trained an RF classifier on 3.1 million labelled sources from Sloan Digital Sky Survey (SDSS) and applied this model on 111 million unlabelled

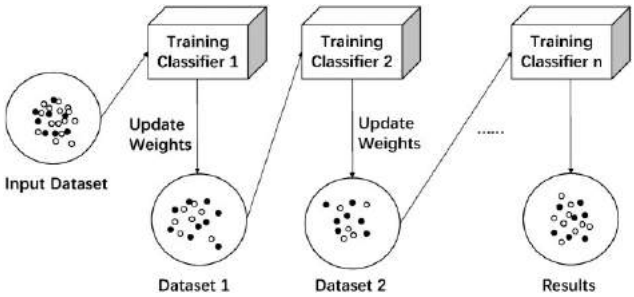

Figure 8. Main idea of boosting. The left data set is regarded as input to train classifier 1 (Training Classifier 1 in Fig. 8). Data set 1: a new data set from training classifier 1 after updating its weights. Train the classifier and update the weights until the classification results converge.

sources. The result showed that the classification probabilities of stars were greater than 0.9 (about 0.99). Besides, RF performed well in the process of searching for rare objects (Hou et al. 2020; Kyritsis et al. 2022). Pattnaik et al. (2021) trained a random forest classifier to determine whether a black hole or a neutron star is hosted by a Low Mass X-ray binaries (LMXBs). It is difficult to accurately classify variable stars into their respective subtypes, hence Pérez-Ortiz et al. (2017) proposed new robust feature sets and used RF to evaluate the classification performance. Akras et al. (2019) used classification tree for identifying symbiotic stars (SySts) from other Hα emitters in photometric surveys. Guo et al. (2022) used random forest to identify white dwarfs in LAMOST DR5. Reis et al. (2018) used an unsupervised random forest to detect outliers on APO Galactic Evolution Experiment (APOGEE) stars. In addition, RF is often compared with other algorithms on classification tasks, and generally, it tends to be better than others (Pérez-Ortiz et al. 2017; Liu et al. 2019; Arsioli & Dedin 2020).

Boosting trains models with adjusted data, that is, the weights of misclassified objects are augmented based on the former models. Fig. 8 is the principle of boosting. Gradient Boosting Decision Tree (GBDT), Adaptive boosting (Adaboost), extreme gradient boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) are prevalent boosting methods. Adaboost is a prominent boosting method that chooses single-layer decision trees as weak classifiers. In each iteration, it trains one weak classifier based on data weights generated in the last iteration. So Adaboost pays more attention on misclassified data. The other essential parameters are weights of each classifier. They are computed based on classification accuracy

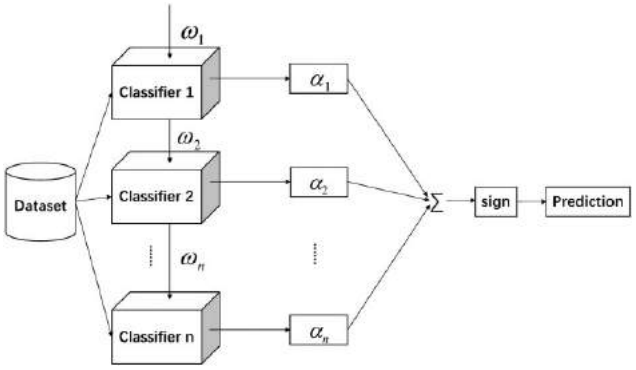

Figure 9. Main idea of adaptive boosting (Adaboost). The left cylinder is input data used to train classifier i (i = 1, 2,....., n). ω_i (i = 1, 2,.....,n) are the weights of data. α_i (i = 1, 2,.....,n) are the weights of classifiers.

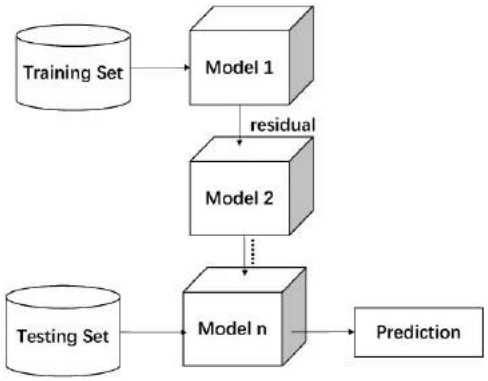

Figure 10. Main idea of gradient boosting decision trees (GBDT). The residual of model i-1 is the input of model i (i = 1,2,.....,n). The goal of GBDT is to make residual as small as possible.

of every classifier. And the final results are obtained after inputting the sum of each weak classifier into a sign function. Fig. 9 is the main principle of Adaboost. GBDT (Pérez-Ortiz et al. 2017; Morice-Atkinson et al. 2018), another typical boosting algorithm, can also be regarded as an optimized version of Adaboost. GBDT chooses the residual from the previous iteration as input to train the next classifier till the residual is close to zero. Besides, GBDT can take more objective functions and train models using negative gradient, whereas Adaboost only sets data weights automatically. Fig. 10 shows the main principle of GBDT. XGBoost optimized GBDT by supporting different meta classifiers, adding regularization to limit model complexity, adapting to different data samplings and so on. Fig. 11 shows the main principle of XGBoost.

GBDT and XGBoost are two powerful ensemble classifiers (Friedman 2001; Chen & Guestrin 2016) and have been applied to spectral classification and rare object identification. Chao, Wen-hui & Ji-ming (2019) used XGBoost to classify star and galaxy on dark sources of SDSS photometric data sets and the results showed that XGBoost outperformed other methods. Hu et al. (2021) searched for Cataclysmic Variables (CVs) in LAMOST-DR7 using LightGBM which is based on the ensemble tree model. They found 225 CV candidates including four new CV candidates which were verified by SIMBAD and published in catalogues. Yue et al. (2021) also identified M sub-dwarfs using XGBoost. In order to get better classification results, many new ensemble algorithms have been

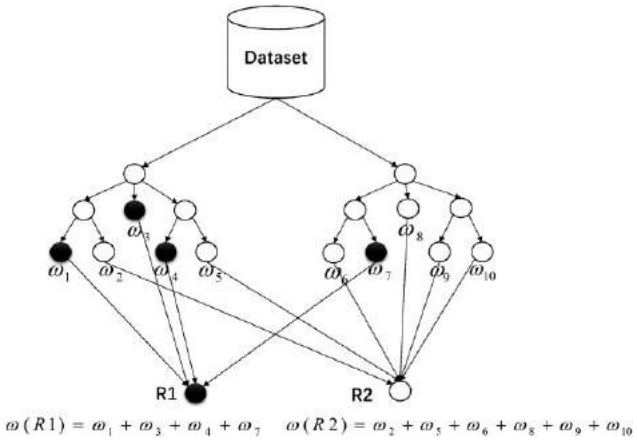

Figure 11. Main idea of XGBoost. ω_i (i = 1, 2,....., n) are the scores of leaves. R1 and R2 are the predicted labels which are the sum of ω_i. The black circles represent data. They are classified as R1 and the white circles are classified as R2. Bigger score between R1 and R2 is the final result.

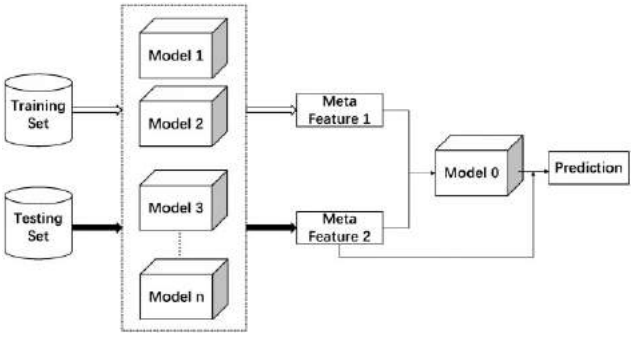

Figure 12. Main idea of stacking. The left cylinders are training sets and test sets. Meta feature 1 and meta feature 2 are the results of model i (i = 1,2,.....,n) on training sets and test sets, respectively. Model 0 is trained by meta feature 1 and meta feature 2. Then meta feature 2 is input into model 0 to predict results.

proposed in recent years (Chao et al. 2020; Chi, Li & Zhao 2022; Zhao, Wei & Jiang 2022).

Stacking uses a new model to fit meta features which are obtained by multipredictors on training sets and testing sets. And this new model will be validated with the following meta features. Fig. 12 introduces the principle of stacking.

Ensemble learning has obtained desirable results in astronomical spectral analysis. And random forest is the most frequently used ensemble method in astronomy. Because it has good generalization performance on large scale high-dimensional data sets. It is good at probabilistic prediction and is insensitive to noise. However, multivalue attribute still troubles RF. In addition, ensemble learning methods are also limited to heterogeneous data, unbalanced data, and optimal parameters (the number of decision tree, weak classifiers).

2.6 Neural network based classification algorithms

Artificial neural network, also known as Multi-Layer Perception (MLP), is a machine learning method that imitates the signal transmission mechanism in the brain. It consists of an input layer, multiple hidden layers, and an output layer. The neural unit in each hidden layer tackles input data and sends results to the next fully connected layer. The output layer generates the final consequences. Fig. 13 is

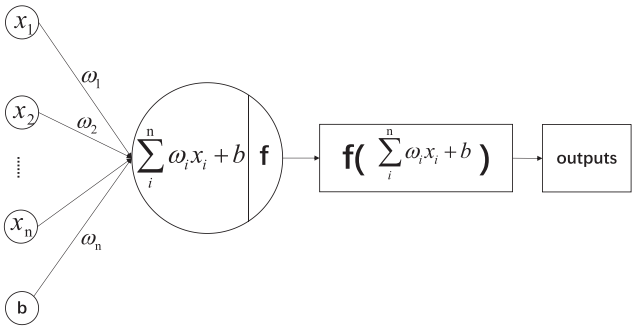

Figure 13. Main idea of neural unit. x_i ($i = 1, 2,\ldots\ldots, n$) is input data. ω_i ($i = 1, 2,\ldots\ldots, n$): weights of x_i. b: (biases) is also the input of neural units. The big circle in the middle contains a linear combination of input and an activation function f.

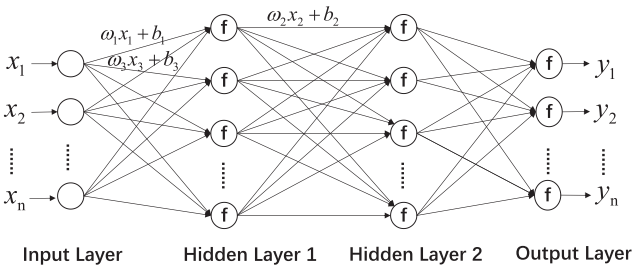

Figure 14. Main idea of artificial neural network. ANN contains multiple fully connected neural units. ω_i ($i = 1, 2,\ldots\ldots, n$) and b_i ($i = 1, 2,\ldots\ldots, n$) are updated during iterations. f is the activation function. y_i ($i = 1, 2,\ldots\ldots, n$) is the final result.

the principle of a neural unit. And Fig. 14 is the main principle of ANN. Particularly, Pseudo-Inverse Learning (PIL) is a classic neural network. It can get globally optimal results and is faster than Back-propagation (BP) algorithm. Besides, it does not require manual tuning of parameters. So it has been used for some simple tasks. However, for complicated tasks, optimal versions of neural network are necessary. Deep learning (DL) is an essential extension of ANN, and it contains more hidden layers and complex network structures (Bergen et al. 2019). Convolutional Neural Network (CNN), Auto Decoder (AE), and Deep Belief Networks (DBN) are three chief methods of DL. Moreover, other variant versions of neural network have been proposed to adapt to different data formats, like Visual Geometry Group (VGG), Residual Networks (ResNet), Recurrent Neural Network (RNN), Generative Adversarial Networks (GAN), and others. Moreover, pre-trained models, attention blocks, transfer learning, and many other tricks have been used to improve the deep learning performance effectively.

Convolutional Neural Network (CNN) consists of convolutional layers that extract image features, pooling layers that reduce dimensionality and fully connected layers that generate results. CNN automatically extracts features without destroying them. So it can get better accuracy and cope with high dimensional data. But its vanishing gradient problem and local optimal phenomenon still annoyed us. Fig. 15 shows the main principle of CNN.

Auto Decoder (AE) is a neural network whose input equals its output and its main idea is sparse code. It restructures the input using an encoder and a decoder. And it has been widely used for noise and dimensionality reduction to visualize data. Fig. 16 is the principle of AE.

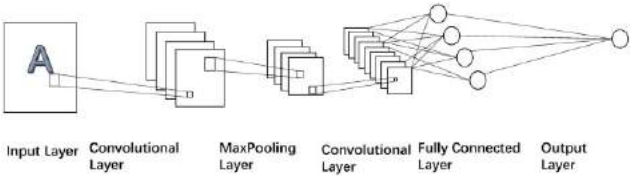

Figure 15. Main idea of convolutional neural network. Convolutional layers are used to learn features from different layers. MaxPooling layer can reduce dimensionality. The role of the fully connected layer is equivalent to the classifier. The output layer is designed to represent the classification results according to the concrete classification task.

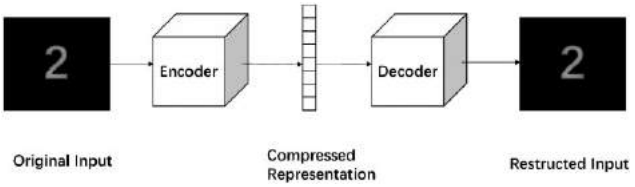

Figure 16. Main idea of auto encoder. Compressed representation of original input is achieved by encoder model. Another vital model is decoder transforming compressed representation into the restructured input.

DBN is a probabilistic generative model. Its generative model builds a joint distribution between observations and labels. DBNs consist of multiple layers of Restricted Boltzmann Machines which is a probabilistic graphical model with stochastic neural network. The output states of each neural unit are activation and deactivation.

Different examples of astrophysical research projects exploiting neural network are listed in Table 5 and summarized in the next parts.

Astronomical spectral classification is a typical task for neural network. Cabayol et al. (2019) used CNN to classify star and galaxy on low-resolution spectra from narrow-band photometry with accuracy over 98 per cent. Jingyi et al. (2018), Astsatryan et al. (2021) used deep CNN to classify quasar and galaxy. Many new improvements of neural network emerged in recent years have been proven to be effective, like residual structures and attention mechanisms (Zou & el al. 2020). A multitask residual neural network was applied to classify M-type star spectra. It reduced the number of parameters in spectral classification and improved the model efficiency (Lu et al. 2020). Compared to other methods, neural network always worked best on the complex data (Aghanim et al. 2015; Guo & Martini 2019; Sharma et al. 2020; Vilavicencio-Arcadia et al. 2020; Chen 2021; Kerby et al. 2021).

Rare object identification is another vital task of neural network (Luo et al. 2008; Guo et al. 2019; Muthukrishna, Parkinson & Tucker 2019; Zou et al. 2019; Jiang et al. 2020; Kou et al. 2020; Margalef-Bentabol et al. 2020; Skoda et al. 2020; Zheng et al. 2020; Tan et al. 2022; Zhang et al. 2022). Shi et al. (2014) searched for metal-poor galaxy (MPG) in large surveys and achieved an MPGs acquisition rate about 96 per cent. Zheng & Qiu (2020) used 1D CNN to search for O stars. Muthukrishna et al. (2019), Fremling et al. (2021), Davison et al. (2022) proposed a software package that used deep learning models to classify the type, age, redshift, and host galaxy of supernova spectra. Qu et al. (2020) identified spectrum J152238.11+333136.1 from LAMOST DR5 and discussed the rare features of P-Cygni profiles.

Neural network based classification algorithms can be used to extract spectral features by different layers (i.e. hidden layers in Fig. 14, convolutional layers in Fig. 15). These layers can automatically learn rich and complex relationships between data. So

Table 5. Investigations of neural network based classification algorithms on astronomical spectra data.

Merits	Caveats	References
High accuracy on stellar spectra and specific spectra[1]	Limited to unbalanced data sets	Cabayol et al. (2019), Shi et al. (2014), Zou & el al. (2020), Farr, Font-Ribera & Pontzen (2020), Wang, Guo & Luo (2017), Liu et al. (2019), Davison, Parkinson & Tucker (2022), Bu et al. (2019), Aghanim et al. (2015), Fuqiang et al. (2014), Arsioli & Dedin (2020), Guo et al. (2019), Rastegarnia et al. (2022), Sharma et al. (2020), Guo & Martini (2019), Fremling et al. (2021), Chen (2021), Flores et al. (2021), Zou, Zhu & Xu (2019), Zheng & Qiu (2020), Tan et al. (2022), Lu, Pan & Yi (2020), Astsatryan et al. (2021), Jingyi et al. (2018), Jiang et al. (2021), Jing-Min et al. (2020), Jiang et al. (2020), Skoda, Podsztavek & Tvrdík (2020), Kerby et al. (2021), Luo et al. (2008), Zheng et al. (2020), Vilavicencio-Arcadia et al. (2020)
Better than other methods[2]	Results are affected by noise	
Generate synthetic data	Poor performance on large redshift objects	
Redshift estimation for quasar, SNe Ia/others	Bad performance on weak features	
Tackle different types of input data sets[3]	Overfitting	
Need less additional information	Misclassification on K/F stars	
Provide vital supplements to categories[4]	High computation time	

Notes. [1] M stars/others, BAL quasars/others, Pulsars/blazars, etc.
[2] RF, template matching, KNN, etc.
[3] Spectra, image, photometric data, etc.
[4] Quasar, star, double-lined spectroscopic binaries, etc.

neural network based algorithms can obtain high accuracy (Moraes, Valiati & Gavião Neto 2013; Fuqiang et al. 2014; Wang et al. 2017; Guo et al. 2019; Liu et al. 2019; Jing-Min et al. 2020; Portillo et al. 2020; Zou & el al. 2020; Jiang et al. 2021). Furthermore, neural network could also handle input features well even without colour or morphological information (Bu et al. 2019; Cabayol et al. 2019) which greatly expanded the size and formats of input data sets.

In short, neural network can learn deep features of data, which will provide subtle differences for classification. More importantly, with the introduction of tricks (i.e. residuals and attention blocks), ANN pays more attention on the valid features. In addition, ANN increases its depth to handle complex and high dimensional data. So it has been widely used in astronomy, such as star/galaxy/quasar classification, MPGs/MRGs classification, rare object identification and spectral feature selection, etc (Rastegarnia et al. 2022). Although neural network model can produce good results, it is a black box that is difficult to interpret results. Compared with decision tree, the results of neural network are difficult for astronomers to analyse the characteristics of celestial objects.

2.7 Gaussian naive Bayes based classification algorithms

Assuming that features are independent, Gaussian naive Bayes based classification algorithms simplify the Bayesian algorithm. They prefer to deal with features in a Gaussian distribution and the maximum posterior probability is the final results. Equation (1) is the objective of Gaussian Naive Bayes based classification algorithms and equation (2) is Gaussian probabilities. Table 6 represents astronomical studies of Gaussian naive Bayes based classification algorithms.

$$y = \underset{c_k}{\operatorname{argmax}}\, P\left(Y = c_k\right) \prod_j P\left(X_j = x_j \mid Y = c_k\right), \tag{1}$$

where

$$P\left(x_i \mid y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \tag{2}$$

δ_y is variance of x_i ($i = 1, 2, \ldots\ldots, n$) and μ_y is average of x_i in equation (2).

Gaussian Naive Bayes based classification algorithms are good at dealing with continuous small data generated from Gaussian distribution. Under the assumption of reliable and sufficient prior spectral information, they could identify rare objects from a large number of spectra data, such as carbon stars (Wallerstein & Knapp 1998; Lloyd Evans 2010; Hoyle et al. 2015; Pruzhinskaya et al. 2019; Arsioli & Dedin 2020). And they were good at reducing noise of stellar spectra, which increased classification accuracy (Kang et al. 2021).

2.8 Logistic regression based classification algorithms

Bayesian Logistic Regression (LR) based classification algorithms obtain posterior probability distributions from linear regression models. And we can get classification results through the sigmoid function. The main researches of LR based classification algorithms are shown in Table 6. Fig. 17 is the principle of Bayesian Logistic Regression based classification algorithms.

LR based classification algorithms can be used for quick regression. However, they cannot get desirable accuracy due to underfitting, bipartition data, and linear data in small feature spaces. In astronomy, logistic regression based classification algorithms were often combined with other techniques to predict physical parameters and classify celestial objects (Luo et al. 2008; Tao et al. 2018; Pérez-Galarce et al. 2021).

2.9 Collaborative representation based classifier and partial least-squares discriminant analysis

Partial least-squares discriminant analysis (PLS-DA) belongs to the discriminant analysis of multivariance data analysis techniques and can be used for classification and discrimination. It handles data in the same cluster rather than data in different clusters. Data in the same group varies widely. And data volumes between groups differ a lot. It extracts principle components of the independent variable X and the controlled variable Y, and finds the relationship

Table 6. Investigations of statistics and ranking on astronomical spectra data.

Merits	Caveats	References
Ranking methods can identify rare objects efficiently	CRC-WPLS is not a prevalent method	Wallerstein & Knapp (1998), Lloyd Evans (2010), Si et al. (2015), Li et al. (2018), Hoyle et al. (2015), Kang, He & Zhang (2021), Du et al. (2016), Daniel et al. (2011), Song et al. (2018), Pérez-Galarce et al. (2021), Tao et al. (2018), Arsioli & Dedin (2020), Pruzhinskaya et al. (2019), Luo et al. (2008)
CRC-WPLS are used on non-linear unbalanced data	Ranking methods also require ample data	

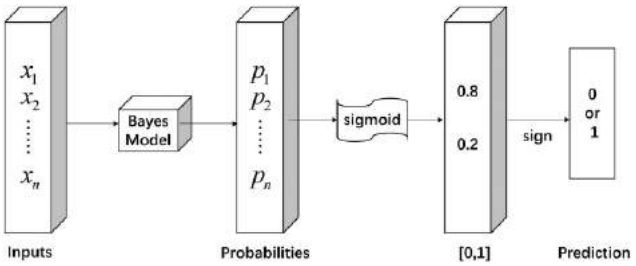

Figure 17. Main idea of Bayesian logistic regression. p_i ($i = 1, 2, \ldots\ldots, n$) are probabilities of x_i ($i = 1, 2, \ldots\ldots, n$) and are generated by Bayes model. Sigmoid function activates p_i ($i = 1, 2, \ldots\ldots, n$) into value between 0 and 1. Sign function transforms probabilities into label 0 or 1.

between principle components in a two high-dimensional space. Table 6 displays the main astronomical researches of CRC-PLS based classification algorithms.

CRC is a novel machine learning algorithm that represents a query by a linear integral of training samples. And CRC classifies the above queries based on the representation (Daniel et al. 2011). It has the ability to handle unbalanced, non-linear, and multilabel data.

CRC-PLS reaps the merits of PLS regression and CRC. So it can classify the high-dimensional spectra data (Song et al. 2018).

2.10 Ranking based classification algorithms

Ranking based positive-unlabelled (PU) learning algorithms have been frequently used in astrophysical object retrieval. Graph based ranking methods successfully identify carbon stars from massive astronomical spectra data, such as manifold algorithm and efficient manifold algorithm (Si et al. 2015), Locally linear embedding. The bipartite ranking is another typical method to improve ranking performance and it has been introduced to search for carbon stars (Du et al. 2016). Alternatively, bagging is a popular method to obtain better performance by integrating different classifiers. The idea of bagging has been well applied in rare object retrieval wonderfully (Du et al. 2016; Li et al. 2018).

The core idea of ranking based classification methods is to learn a ranking based model which usually ranks data sets by pre-defined evaluation methods. They have two goals: (1) positive samples are ranked ahead of negative samples. (2) the scores of related samples tend to be similar. Many optimal ranking methods have emerged to improve classification performance and reduce time consumption, such as efficient manifold algorithms and bagging TopPush. And these methods have already discovered carbon stars from extensive spectra data which is a significant supplement to the catalogues of carbon stars (Table 6).

3 EXPERIMENT ANALYSIS

Recently, lots of basic or improved classification algorithms have been successfully applied to various astronomical data analyses. However, due to the diversity of classification tasks and classification data, it is difficult to assess the advantages and disadvantages of these methods from the current literature. So, in this section, we construct unified experimental spectral data sets from LAMOST survey and SDSS survey to evaluate the commonly used methods.

3.1 Experimental data introduction

In the experimental design, we construct several groups of data sets using the spectra data from LAMOST (Luo et al. 2015) and SDSS.

LAMOST (The Large Sky Area Multi-Object Fiber Spectroscopic Telescope, also known as Guo Shou Jing Telescope) is a special reflective Schmidt telescope with an effective aperture of 3.6–4.9 m and a field of view of 5°. It is equipped with 4000 fibres, a spectra resolution of $R \approx 1800$, and a wavelength ranging from 3800 to 9000 Å (http://www.lamost.org/public/?locale = en). Its scientific goal is to make a 20 000 deg^2 spectroscopic survey (DEC: $-10^\circ \sim +90^\circ$). After seven years of surveying, LAMOST has observed tens of millions of low-resolution spectra data, providing important data for astronomical statistical research.

The Sloan Digital Sky Survey (SDSS) is an international collaboration of scientists to build the most detailed 3D imagery of the Universe. It uses a wide-field telescope with a diameter of 2.5 m and a field of view of 3°. The photometric system is matched with five filters in u, g, r, i, and z bands to photograph celestial objects. It covers 7500 deg^2 of the sky around the South Galactic Pole and records data on nearly 2 million celestial objects.

Experimental data are selected from LAMOST DR8 and SDSS DR16. The LAMOST DR8 data sets include a total of 17.23 million released spectra. The number of high-quality spectra of DR8 (that is, the S/N > 10) reaches 13.28 million and DR8 includes a catalogue of about 7.75 million groups of stellar spectral parameters. The SDSS DR16 covers more than one-third of the sky and contains about 5789 200 total spectra and 4846 156 useful spectra. And DR16 contains new optical and infrared spectra, including the first infrared spectra observed by Las Campanas Observatory in Chile.

We select and pre-process the spectra from four aspects. These are shown in Table 7.

(1) Data release. We select spectra from LAMOST DR8 and SDSS DR16.

(2) Extinction problem. In order to decrease the influence of reddening on classification performance, 1D spectra in data sets are selected from LAMOST ($45^\circ < l$) (Yang et al. 2022a).

(3) Flux calibration. LAMOST uses relative flux calibration. We cut off the overlapping region (5700 Å$<\lambda <$ 5900 Å) known to have calibration issues to minimize their effect on our classification.

Table 7. Data preprocessing.

	Data selection and preprocessing
Data release	LAMOST DR8, SDSS DR16
Extinction	1D spectra from LAMOST($l > 45°$)
Redshift	Rest wavelength frame spectra for star/galaxy/quasar
Flux calibration	Relative flux calibration: cut off 5700 Å–5900 Å

(4) Redshift. For star/galaxy/quasar classification, we convert original spectra into the rest-frame wavelengths by applying the redshift values from LAMOST and compare the performance of classification on the rest wavelength frame spectra and original spectra. Because the radial velocity of stellar spectra are small under the current resolution of LAMOST, which has little influence on classification results. Spectra for stellar classification are left in the observed frame wavelengths.

We determine three classification tasks among multiple astronomical researches, including A/F/G/K stars classification, star/galaxy/quasar classification, and rare object identification. Rare objects includes carbon stars (Wallerstein & Knapp 1998; Lloyd Evans 2010; Gigoyan et al. 2012), double stars, artefacts: bad merging of red and blue segments (A common phenomenon that occurs in the spectra of LAMOST).

We design six groups of data sets for the above tasks. Data sets 1–data sets 3 are constructed for A/F/G/K stars classifications. They are divided by data characteristics, S/Ns and data volumes, and each data set contains three or four sub-data sets. Datasets 4 are used to evaluate the classification performance of star/galaxy/quasar on original spectra and rest wavelength frame spectra. Data set 5 is used to identify rare objects: carbon stars, double stars, and artefacts. And the classifier is trained on 200 rare objects and 19 900 other non-rare objects. Non-rare objects include 10 000 normal stars, 6500 galaxies, and 3400 quasars. We analyse the results of rare object identification by accuracy, precision, recall, and F1 score. Spectra of the first five groups of data sets are selected from LAMOST. Because the sources of LAMOST have considerable overlaps with SDSS, we construct the matching data sets (data sets 6 in Table 8) from SDSS and LAMOST to compare the classification performances on them. The analyses of experimental results on data sets 6 are elucidated in Section 3.2.1.

The composition of testing sets in all data sets is the same as their training sets. The ratio of training sets and testing sets for data sets 1, data sets 2, data sets 3, data sets 4, and data sets 6 is 8:2 and the ratio of training sets and testing sets for data set 5 is 1:1. Details of data sets are shown in Table 8.

3.2 Result analysis

In this section, nine basic methods including K-Nearest Neighbour, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting Decision Tree, Logistic Regression, Pseudo Inverse Learning, and Convolutional Neural Network are tested on astronomical spectra data and we fairly evaluate the classification performance.

Our experiments use grid search (Syarif; Prügel-Bennett & Wills 2016) to identify the optimal parameters of each algorithm. And we take the average accuracy of 5-fold cross validation (Fushiki 2011) as the final accuracy to avoid the influence of sample selection.

3.2.1 Performance analysis on 1D spectra, PCA, and line indices

Fig. 18 represents the accuracy of nine basic algorithms on three data characteristics (1D spectra, PCA, line indices).

In the classification on 1D spectra, CNN achieves the highest accuracy. Because it can extract complex features through different layers. However, CNN still suffers from two unavoidable drawbacks. One is that it has to spend a long time to obtain the optimal model. The other is overfitting which cannot be easily eliminated even by L2 regularization or dropout method. In order to reduce the training time, we can extract features by PCA and classify the pre-processed spectra. Because accuracy on PCA features is equal to that on 1D spectra and the training time is shorter.

In Fig. 19, A stars and K stars can be distinguished admirably whereas F stars and G stars have disappointing accuracy. Because F stars and G stars are more similar than A stars and K stars in the global shape of 1D spectra. Stellar rotation might become another reason for the misclassification because it broadens spectral lines and might cause the global shape of 1D spectra if lines are blended because of insufficient spectral resolution. So it is necessary to alleviate the influence of stellar rotation on classification. Moreover, researchers can use other spectra characteristics to avoid the caveats of 1D spectra. Results also show that LR, Pseudo Inverse Learning (PIL),

Table 8. Data sets of spectral classification.

	Data sets introduction[1]	Data components[2]	S/N	Characteristics
Data sets 1	A/F/G/K stars classifications on four characteristics	A : F : G : K Stars = 5000 : 5000 : 5000 : 5000	>10	1D Spectra
		A : F : G : K Stars = 5000 : 5000 : 5000 : 5000	>10	PCA (100 dimensions)
		A : F : G : K Stars = 5000 : 5000 : 5000 : 5000	>10	Line Indices
Data sets 2	A/F/G/K stars classifications on three S/Ns	A : F : G : K Stars = 5000 : 5000 : 5000 : 5000	<10	1D Spectra
		A : F : G : K Stars = 5000 : 5000 : 5000 : 5000	10-30	1D Spectra
		A : F : G : K Stars = 5000 : 5000 : 5000 : 5000	>30	1D Spectra
Data sets 3	A/F/G/K stars classifications on four volumes	A : F : G : K Stars = 2000 : 2000 : 2000 : 2000	>10	1D Spectra
		A : F : G : K Stars = 5000 : 5000 : 5000 : 5000	>10	1D Spectra
		A : F : G : K Stars = 10000 : 10000 : 10000 : 10000	>10	1D Spectra
		A : F : G : K Stars = 20000 : 20000 : 20000 : 20000	>10	1D Spectra
Data sets 4	Star/galaxy/quasar classifications	star : galaxy : quasar = 5000 : 5000 : 5000	stars: >10,	Original Spectra
		star : galaxy : quasar = 1000 : 1000 : 1000	galaxies, quasars: all	Rest Wavelength Frame Spectra
Data set 5	Search for rare objects [3]	rare objects : normal stars : galaxies : quasars = 200 : 10000 : 6500 : 3400	normal stars : >10, galaxies, quasars : all	1D Spectra
Data sets 6	A/F/G/K stars classifications on LAMOST and SDSS	A : F : G : K = 5824 : 5380 : 4151 : 6240(LAMOST)	>10	1D Spectra
		A : F : G : K = 5797 : 5355 : 4144 : 6229(SDSS)		

Notes. [1] Spectra of data sets 1–data set 5 are selected from LAMOST. Spectra of data sets 6 are selected from LAMOST and SDSS.
[2] The values of the data components in this table are the actual data volume.
[3] Rare objects: carbon stars, double stars, artefacts : bad merging of red, and blue segments.

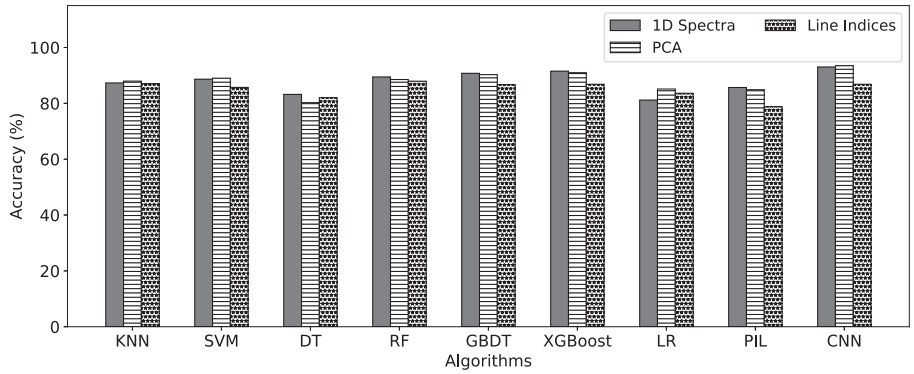

Figure 18. Accuracy of algorithms on different data characteristics. Three different types of bars stand for various data characteristics.

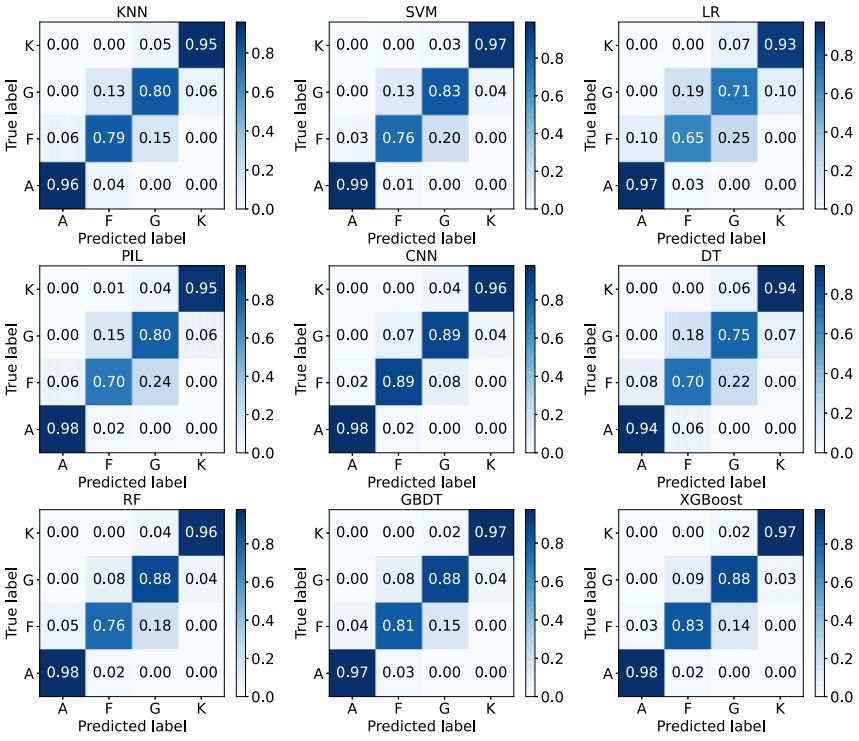

Figure 19. Accuracy of algorithms on 1D spectra of A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

DT cannot get desirable results on F stars and G stars due to the weak spectral shapes. While strong classifiers (CNN, ensemble methods, SVM) show superiority.

PCA is a useful dimensionality reduction tool in many fields. Technically, it extracts principle components of spectra. And the principle components preserve the main information of spectra as much as possible. So accuracy shows little difference with 1D spectra (Figs 18–20). However, the consistent results cannot be explained well because linear PCA may be misleading to tackle the non-linear spectral lines. This phenomenon has also confused researchers (Tao et al. 2018). And spectra pre-processed by PCA are a linear sum of different dimensional characteristics from 1D spectra which lacks concrete (astro)physical meaning. These problems need to be explored in the future. The main merit of PCA is that the spectra pre-processed by PCA can reduce the number of features and the computation time. So it has been widely used in astronomical tasks.

Line indices are vital features for spectral analysis. They refer to the relative intensity of absorption or emission lines produced by certain elements. And stellar absorption lines can be used to distinguish stars. Fig. 18 illustrates the results of nine basic algorithms on line indices. Overall, nine basic classification algorithms performed similarly. Compared with the results on 1D spectra, simple KNN is superior to CNN in the low dimensional space of line indices. Because the powerful feature selection of CNN tends to show advantages in high dimensional space. Fig. 21 show more misclassifications between A stars and F stars. Misclassification between F stars and G stars has decreased a little. And we can clearly see that F stars can be distinguished better than other stars.

Comparative results analysis of LAMOST and SDSS. As can be seen from Fig. 22, the classification algorithms perform better on SDSS instead of LAMOST. The reason may be that the calibration quality of LAMOST will be influenced by fibre-to-fibre sensitivity

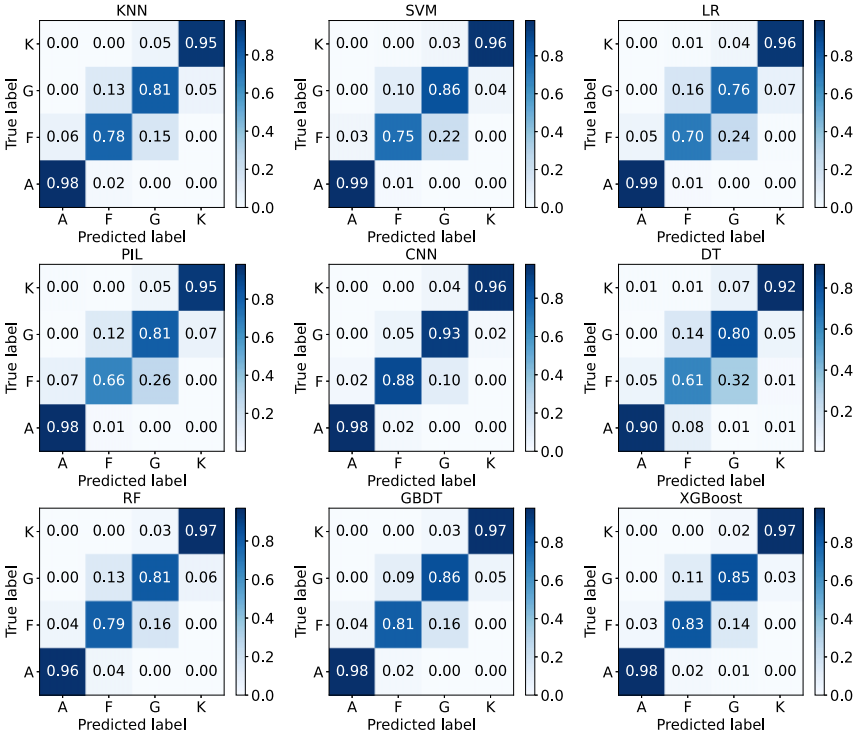

Figure 20. Accuracy of algorithms on PCA of A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

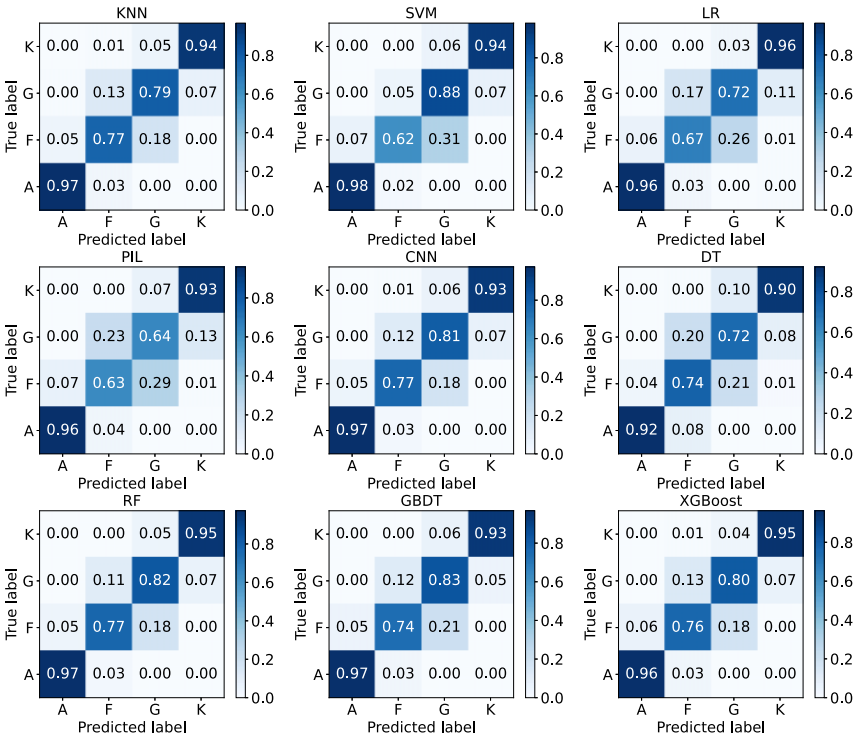

Figure 21. Accuracy of algorithms on line indices of A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

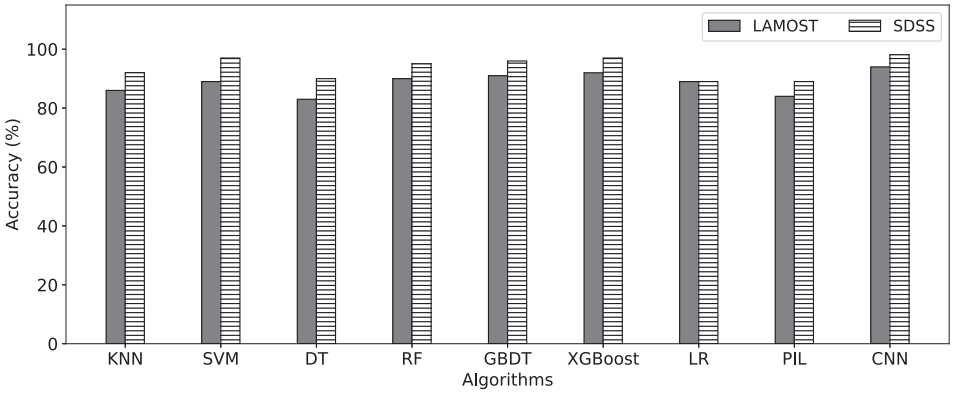

Figure 22. Accuracy of algorithms on A/F/G/K stars of LAMOST and SDSS. Two different bars represent the spectra from LAMOST and SDSS.

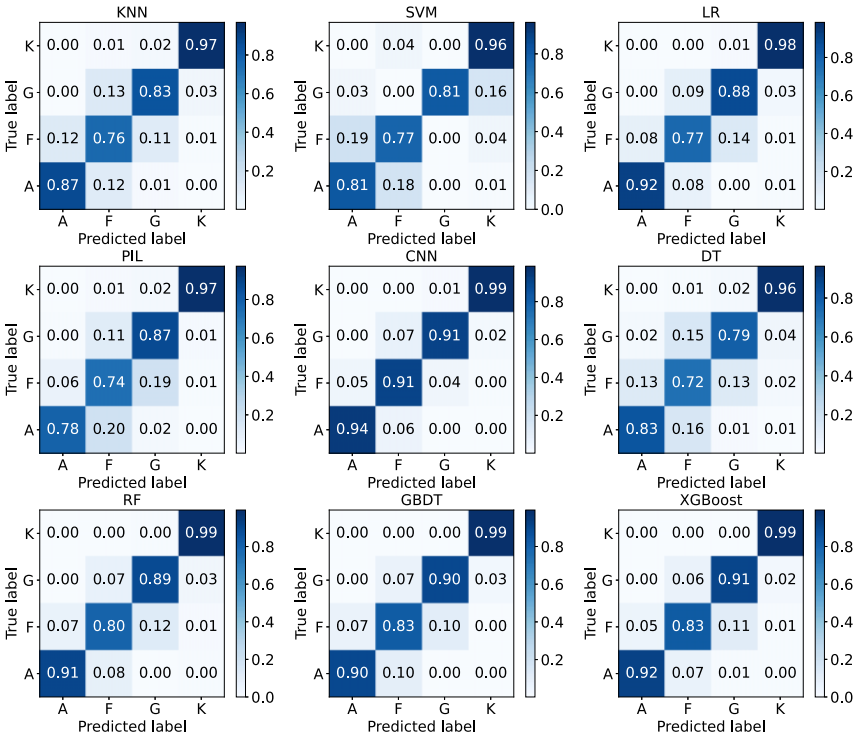

Figure 23. Confusion matrices of algorithms on A/F/G/K stars of LAMOST. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

variations, further causing the slight differences on classification results. As shown in Fig. 23, all classification algorithms perform best on K-type stars from LAMOST. But they perform poorly on F-type stars from LAMOST. Similarly, the performance of classification algorithms on F-type stars from SDSS is bad (Fig. 24). And it can be clearly seen that the performance of classification algorithms on A, G, K stars from SDSS is similar, but slight better than that from LAMOST.

3.2.2 Performance analysis on spectra qualities

On the whole, the accuracy is in direct proportion to S/N (Fig. 25). Paying more attention on S/N>30, we can draw a conclusion that SVM, ensemble methods, and CNN can achieve better results than KNN. And, the classification performance of PIL is better than that of LR. Because PIL can extract complicated features through a simple three-layer neural network while LR fails in high-dimensional space.

The accuracy of classification on S/N: 10–30 drops completely because spectral data on S/N: 10–30 are always mixed with noise. CNN continues to remain top of the nine basic algorithms because it has added regularization and dropout methods to alleviate overfitting.

It is difficult to mine information from spectra on S/N<10 which are often regarded as unqualified spectra. As a result, it is prevalent to obtain low accuracy on spectra with S/N<10. We divide algorithms into three parts according to their classification accuracy. Obviously, SVM, CNN, ensemble methods, and LR are the leading echelons followed by PIL. SVM shows robustness on S/N<10. Because the soft margin of SVM guarantees that most spectra are classified correctly even for some misclassified samples. Likewise, CNN gains 70 per cent accuracy depending on the strong ability of feature selection. GBDT and XGBoost adopt gradient boosting methods to

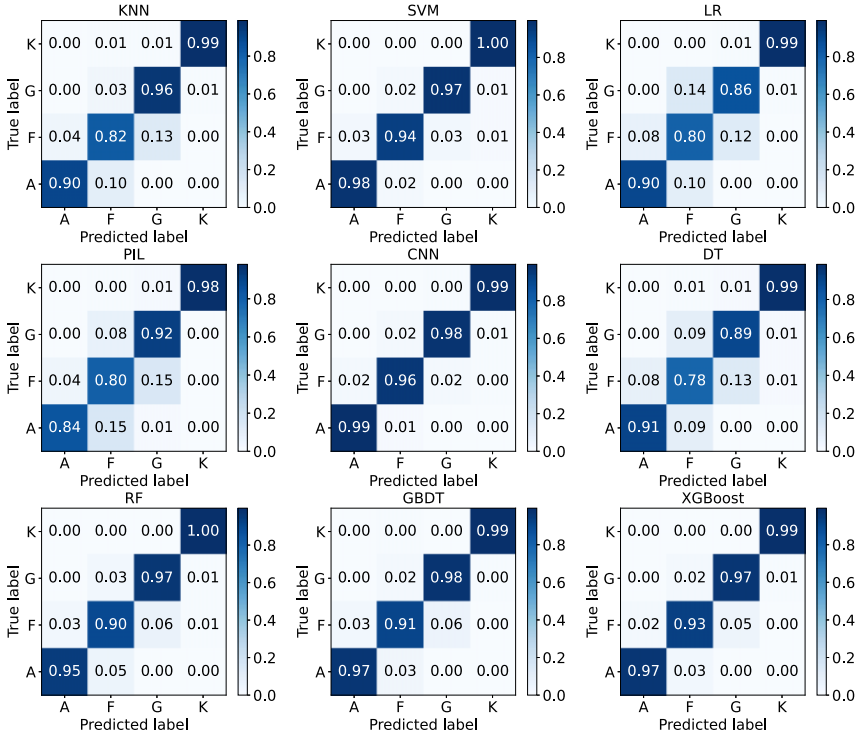

Figure 24. Confusion matrices of algorithms on A/F/G/K stars of SDSS. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

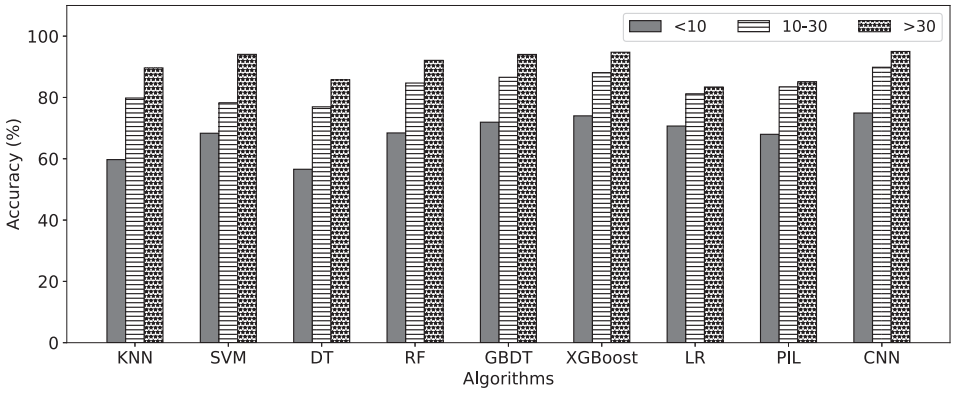

Figure 25. Accuracy of algorithms on different S/Ns. Different colour bars stand for different S/Ns.

reduce errors and attain higher accuracy than RF which only merges different decision trees. KNN and decision tree cannot satisfy us. They perform worst. Because KNN uses Euclidean distance as a distance metric. So it is susceptible to noise. Likewise, decision tree cannot find proper splitting features because of noise. We can find misclassification of spectra on S/N < 10 from Figs 26, 27 and 28, such as the poor performance of PIL and decision tree methods on F stars.

3.2.3 Performance analysis on different data volumes

Figs 29 and 30 show the performance of nine basic classification algorithms on the four different data volumes.

There is a slight improvement in the accuracy with the increase of data volumes. Because the large number of spectral data will provide more information to obtain better classifiers.

Fig. 31 shows the computation time of nine basic classification algorithms on four different data volumes. Compared with other algorithms, SVM and CNN spend more time on classification. Besides, the computation time of SVM, CNN, and LR increases rapidly as the data volumes increase.

There is little difference in the confusion matrices of different data volumes. And the main misclassification exists between F stars and G stars in Figs 32–35.

3.2.4 Performance analysis of star, galaxy, and quasar classification

As can be seen from Fig. 36, most classification algorithms perform better on the rest wavelength frame spectra than on the original spectra. Because redshift causes feature shift problems, overlapping

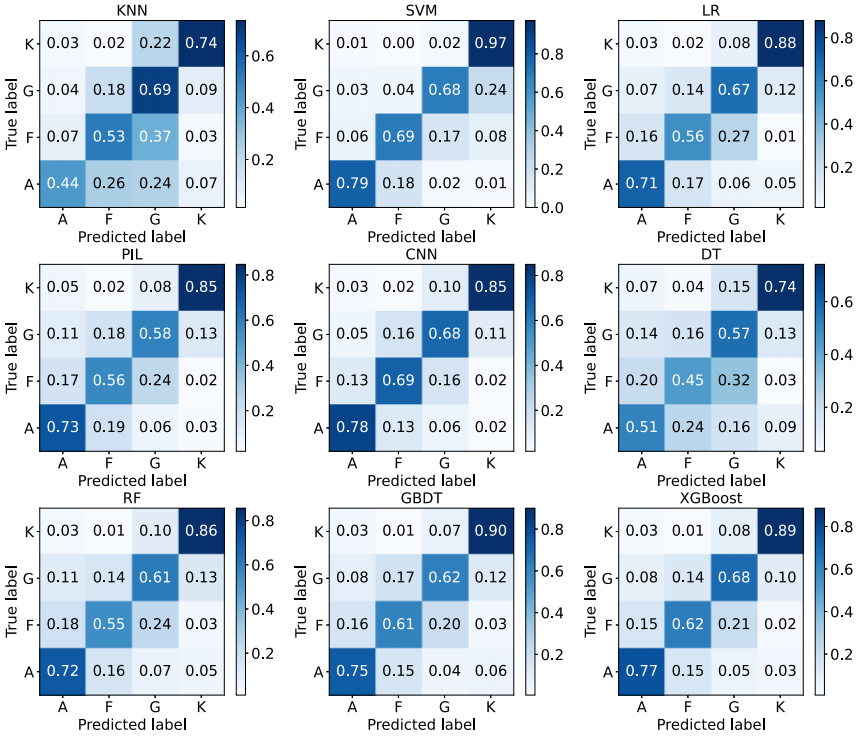

Figure 26. Accuracy of algorithms on S/N<10 of A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

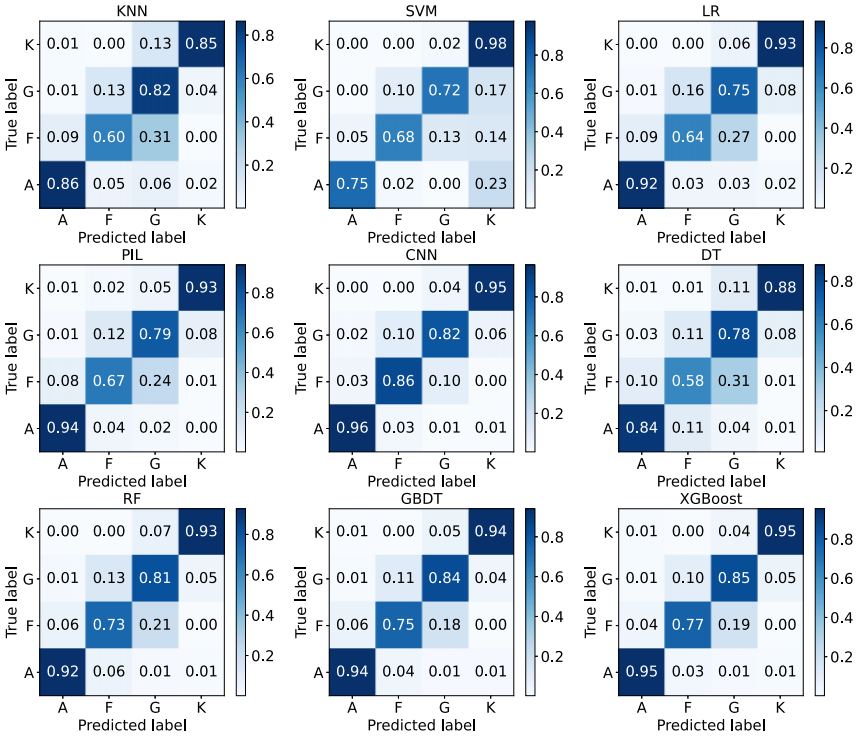

Figure 27. Accuracy of algorithms on S/N: 10–30 of A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

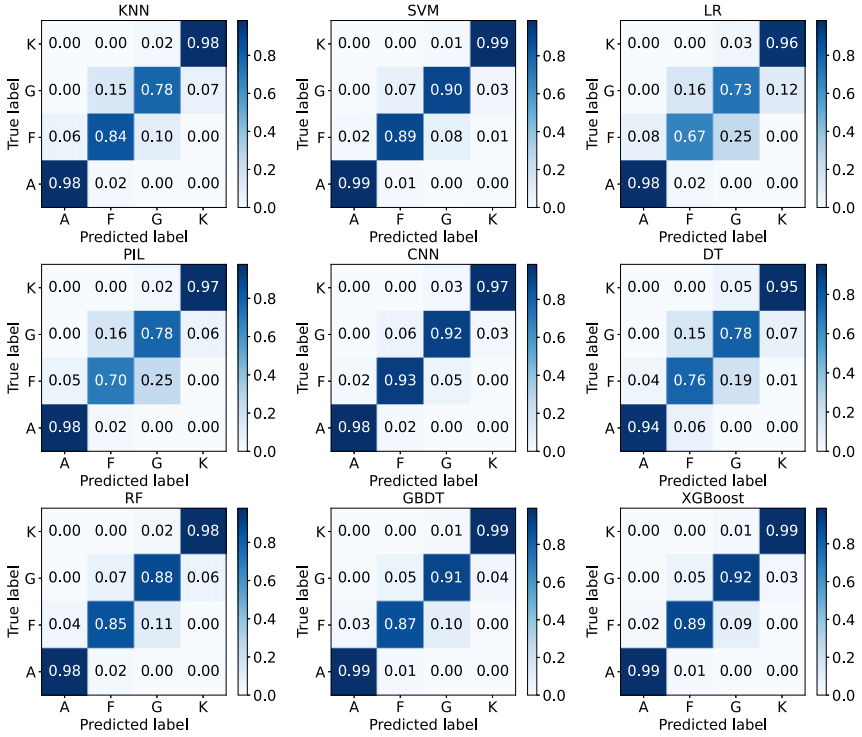

Figure 28. Accuracy of algorithms on S/N>30 of A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

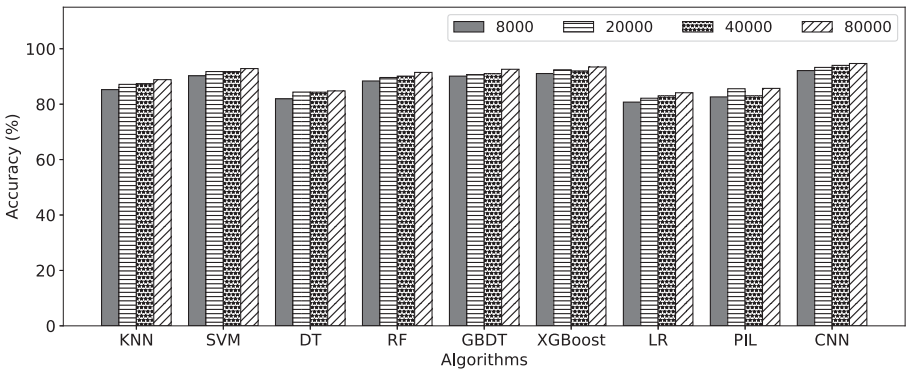

Figure 29. Accuracy of algorithms on different data volumes. Four different bars stand for four types of data volumes.

phenomena between nearby galaxies and high-velocity stars. These problems affect the performance of classification algorithms on original spectra. We also found that LR, PIL, and CNN algorithms perform better on original spectra. Because LR fits more complex polynomials to classify spectra and the other two methods learn deep features for better classification. So the above issues caused by redshift make little influence on the classification performance of these methods. In addition, the dimensionality of rest wavelength frame spectra is reduced and some information will be lost, which will also lead to poor classification performances of LR, PIL, and CNN.

Pay more attention on the classification algorithms in Fig. 36, they can be divided into three parts: CNN, SVM, RF, GBDT, XGBoost; DT, LR, PIL; KNN. CNN performed better than others for its powerful ability of feature selection. The classical classifier SVM can also find a suitable hyperplane to separate the rest wavelength frame spectra. Methods such as RF, GBDT, XGBoost can classify rest wavelength frame spectra well due to their integration. Decision tree and random forest cannot choose the split nodes well because of the inconsistent features. KNN cannot classify galaxy and quasar well. Because the feature lines are inconsistent on spectra shape and position due to redshift. Misclassification can also be found in Figs 37 and 38.

3.2.5 Performance analysis on rare targets

Compared with the classifications performance of A/F/G/K stars classification on 1D spectra, the classification algorithms perform bad when searching for carbon stars, double stars, and identifying artefacts (Figs 39–41). Because the imbalanced data sets have a bad impact on the classification performance.

Due to the obvious characteristics of carbon stars, classification performance of carbon stars is better than that of double stars and

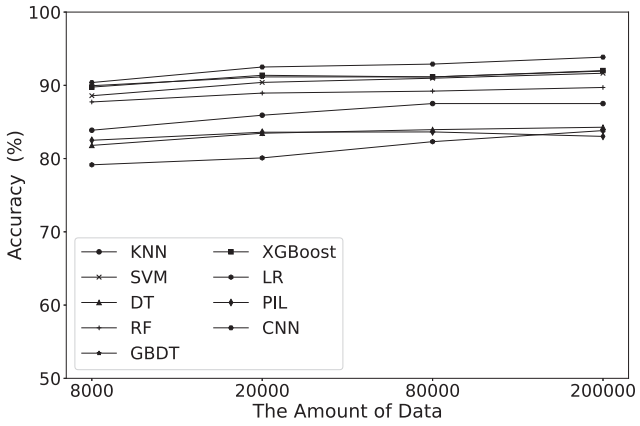

Figure 30. Accuracy of algorithms on different data volumes. Different shapes in lines represent different classification algorithms.

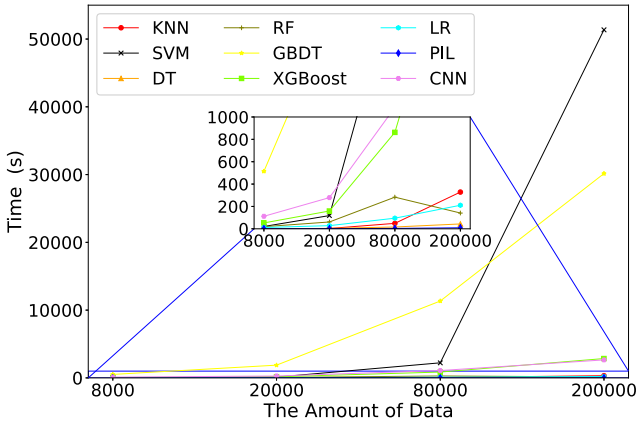

Figure 31. Time of algorithms on different data volumes. Different line colours mean various algorithms. Time between 0 and 1000 is clearly shown in the middle rectangle.

artefacts. As can be seen from Fig. 39, the classic binary classifier SVM and ensemble learning methods (RF, GBDT, XGBoost) perform better than other algorithms.

The classification algorithms have the worst performance in identifying double stars due to the mutual interference of the overlapping parts in double stars. This problem will affect the classification performance on carbon stars. Precision in Fig. 40 shows that the ensemble learning (RF, GBDT, XGBoost) can identify some double stars accurately. But the recall in Fig. 40 shows that a large number of double stars will be missed.

It can be seen from Fig. 41 that classification performance of identifying artefacts is between the carbon stars and double stars. Several ensemble algorithms can also find these rare stars accurately. Compared with the double stars, the recall rate of artefacts is relatively improved. It means that several integration algorithms and KNN can identify more artefacts. But it is inevitable that many artefacts will be missed.

4 SOURCE CODE AND MANUAL

Source codes used in this paper are provided on https://github.com/shichenhui/SpectraClassification. Algorithms in the code category are shown in Table 9. Because the parameters of algorithms have a significant impact on the classification results, we also provide the parameters of algorithms on each data set and the parameters are optimized by grid-search method provided by SKLEARN package.

The codes are written in python which is widely used for machine learning and data analysis. Dependent packages of our codes include NUMPY (Harris et al. 2020), SKLEARN, MATPLOTLIB (Hunter 2007), PANDAS, SCIPY. Each algorithm is organized by the following steps: (1) load training data sets and testing data sets; (2) configure the parameters of classification models; (3) train models on the training data sets; (4) classify the testing data sets by training models; (5) evaluate the performance of training models. To avoid the influence of sample selection on the training data sets, we use 5-fold cross-validation to split data sets and evaluate models. But this is not necessary for practical applications.

These codes load data from ∗.csv files which store tabular data in the form of text. And a row of data is a spectrum. You need to convert your spectra data into this format or modify the data loading mode. Some basic algorithms are directly implemented from SKLEARN packages.

The parameter K of KNN is not a fixed value (default value in SKLEARN is 5). Generally, a smaller value is often selected according to the sample distributions. And an appropriate K value can be selected by cross-validation. Besides, it adopts Euclidean distance as distance metrics to get good results in low dimensional space. Other distance metrics can also be applied in KNN to avoid the disadvantage of Euclidean distance.

SVM needs to select kernel functions. There are many kernel functions: linear kernel function, polygon kernel function, RBF kernel function, sigmoid kernel function, etc. The current improvement of SVM is combined with other methods to classify the large-scale data sets.

Feature selection criteria and feature splitting criteria are two important parameters of decision tree. Different feature selection methods (information entropy, information gain, Gini index) correspond to different decision trees. Features splitting parameters can be 'best' or 'random'. The former is to find the optimal division point from all division points of the features, the latter is to find the local optimal division point from the randomly selected division points. Generally, 'best' is often used for the small number of samples and 'random' for the large number of samples. Other parameters like tree depth and the number of trees are also needed to be determined.

Ensemble learning algorithms (i.e. random forest, GBDT, and XGBoost) are integrated by decision trees. We need to choose the number of integrated trees. Methods in SKLEARN use 100 decision trees by default. But GBDT cannot be parallel, we need to reduce the number of decision tree appropriately. Other parameters in decision tree can be set up according to the introduction in the previous paragraph.

Logistics regression is a binary classifier. It integrates multiple LR classifiers for multiclassification tasks. The integration strategy is always 'OVR'. And it uses 'L1' and 'L2' regularization to reduce overfitting. 'L2' is more commonly used. But for high dimensional data, 'L1' penalty can help you reduce the impact of unimportant features.

The good design of neural network structures is important for ANN based methods. We find that 1D convolutional structure can extract spectral features well. So for spectral classification, the performance of CNN is better than that of fully connected neural network. There are many layers in computer vision. But for data in the format of vector, we do not need to stack too many layers in the neural network structures. Likewise, 'L1' and 'L2' regularization can be used to reduce overfitting.

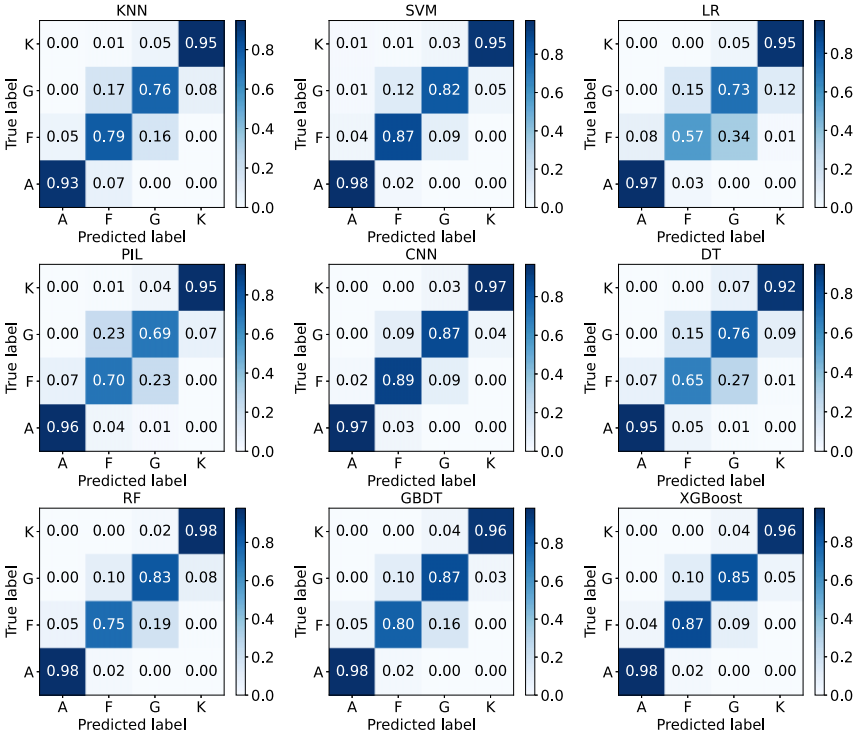

Figure 32. Accuracy of algorithms on data volume of 8000 for A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

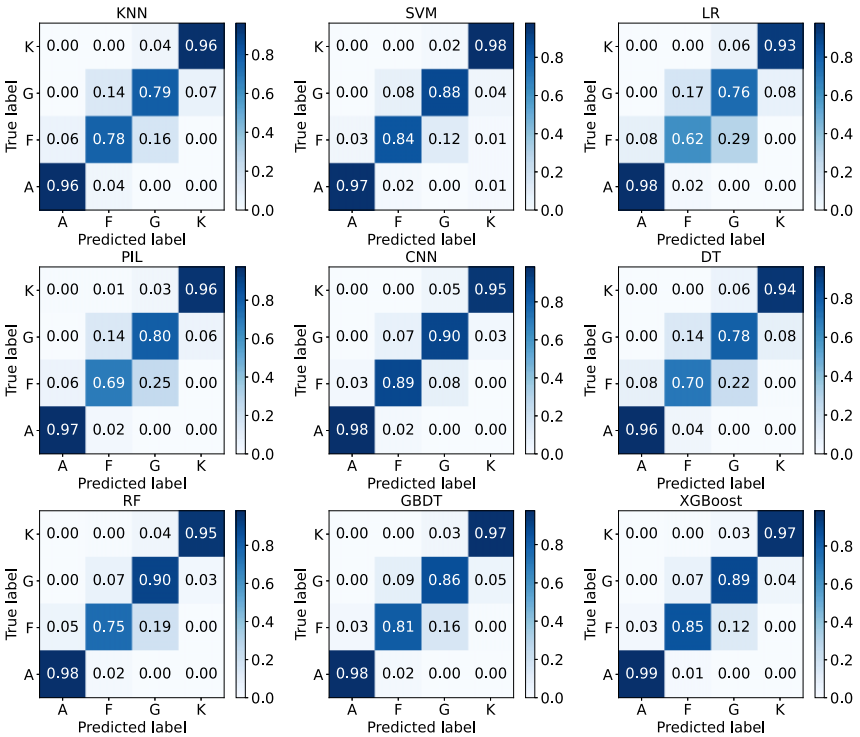

Figure 33. Accuracy of algorithms on data volume of 20 000 for A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

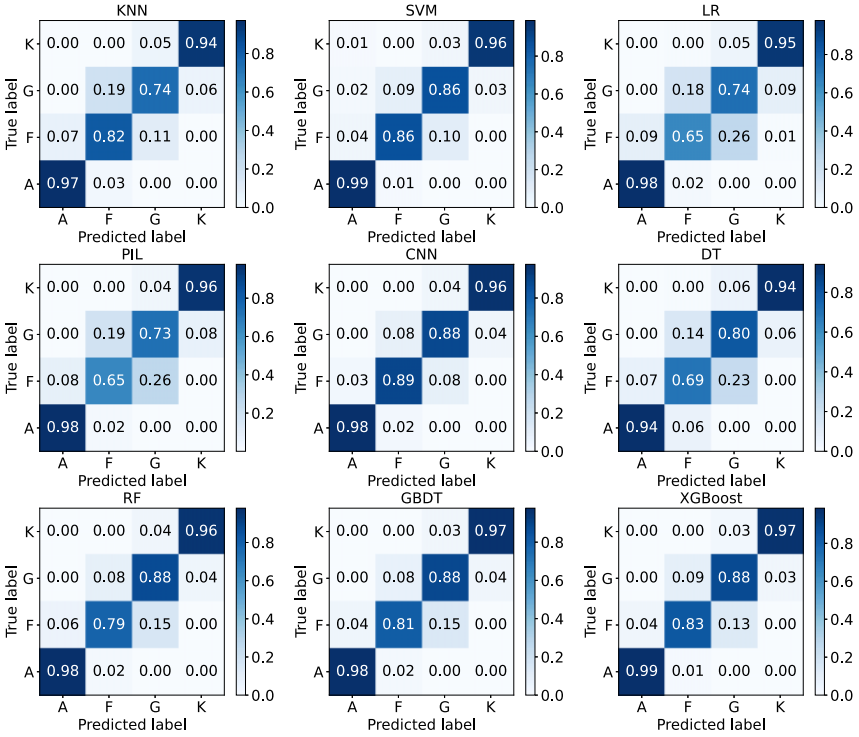

Figure 34. Accuracy of algorithms on data volume of 40 000 for A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

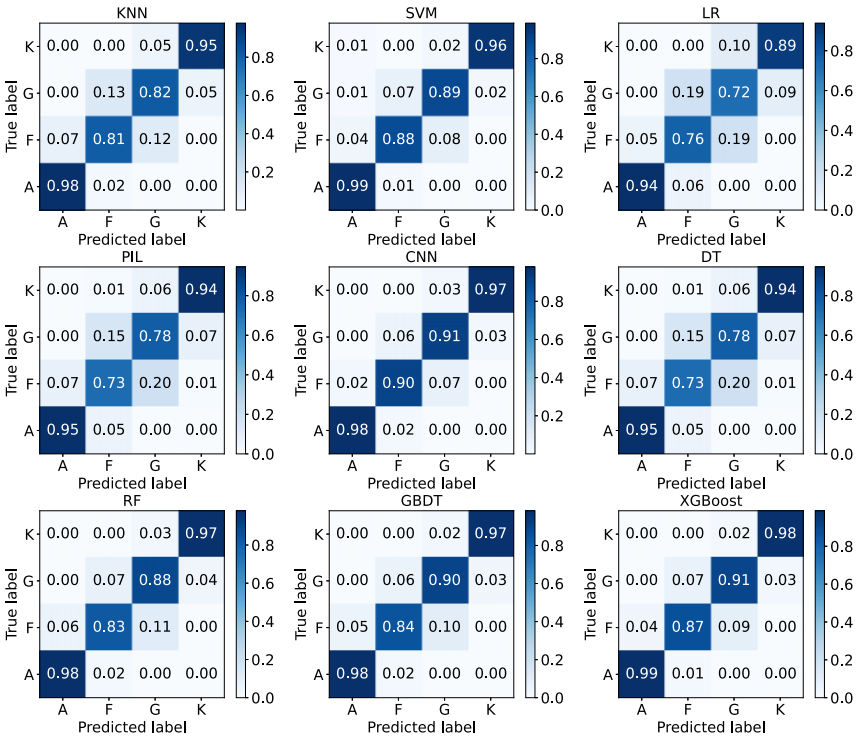

Figure 35. Accuracy of algorithms on data volume of 80 000 for A/F/G/K stars. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

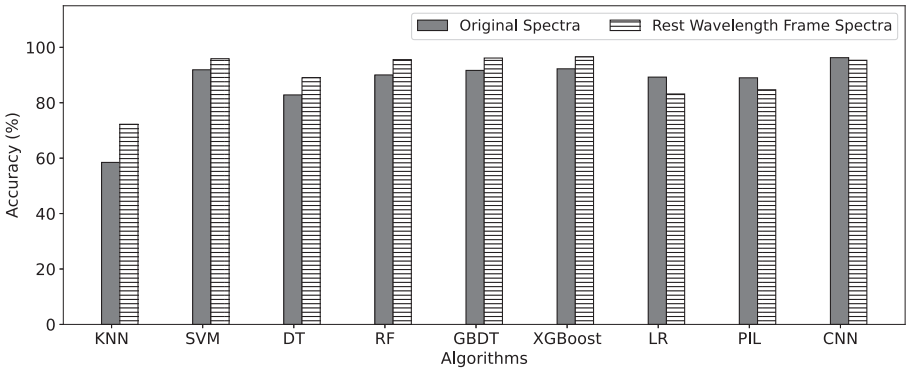

Figure 36. Accuracy of algorithms on star/galaxy/quasar on original spectra and rest wavelength frame spectra. Two bars represent two spectra characteristics.

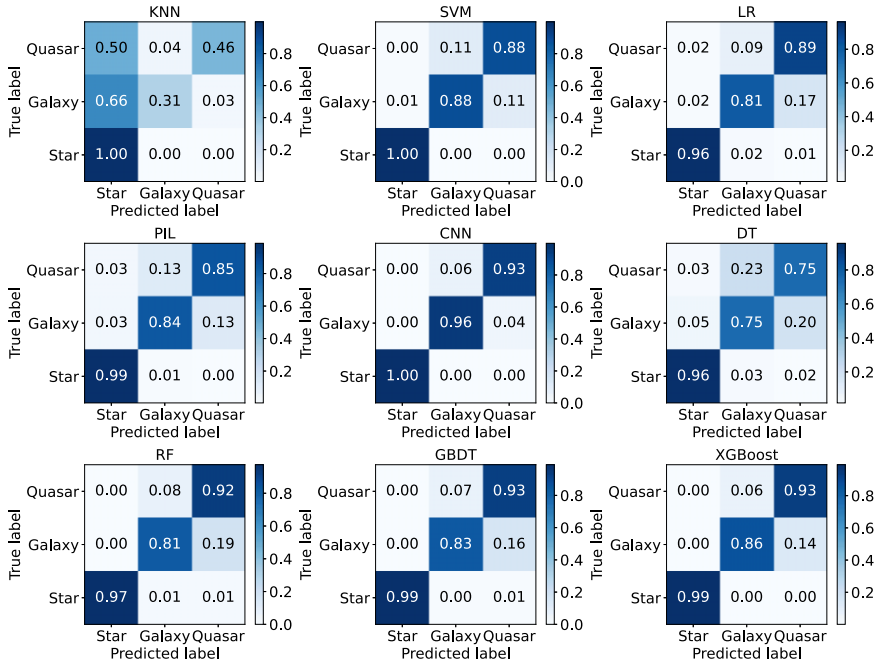

Figure 37. Accuracy of algorithms on star/galaxy/quasar on original spectra. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

5 DISCUSSION

In this paper, we investigate the classification methods used for astronomical spectra data. We introduce the main ideas, advantages, caveats, and applications of classification methods. And data sets are designed by data characteristics, data qualities, and data volumes. Besides, we experiment with nine basic algorithms (KNN, SVM, LR, PIL, CNN, DT, RF, GBDT, XGBoost) on A/F/G/K stars classification, star/galaxy/quasar classification, and rare object identification. Experiments on data characteristics also include the comparative experiments on the matching sources from the LAMOST survey and SDSS survey.

For A/F/G/K stars classification, the accuracy on 1D spectra and PCA shows little difference while PCA spends less time in the training stage. Because it reduces the spectra dimensionality. So PCA is often used to classify large-scale and high dimensional data sets. Among nine basic methods, CNN performs best on 1D spectra and PCA, due to its powerful ability for feature selection. For the classification on line indices, KNN shows superiority among other methods. The performance of classification on SDSS is better than that on LAMOST. Because the calibration quality of LAMOST is undesirable, which is affected by many factors (i.e. fibre-to-fibre sensitivity variations). In addition, high-quality spectra and a large number of samples help us to train models. But with the growth of data volumes, the training time of some models will also increase greatly. So it is necessary to improve the classification speed on large-scale data sets.

As for star/galaxy/quasar classification, most performance of classification on rest wavelength frame spectra is better than that on original spectra. Because redshift causes feature movement on original spectra. But for some algorithms (PIL, LR, CNN), the performance of classification on the original spectra is better than that on the rest wavelength frame spectra. Because original spectra have much information. These methods can extract feature well and are less influenced by redshift. For this task, SVM which is good at binary classification and CNN with powerful ability for feature selection perform better than other methods.

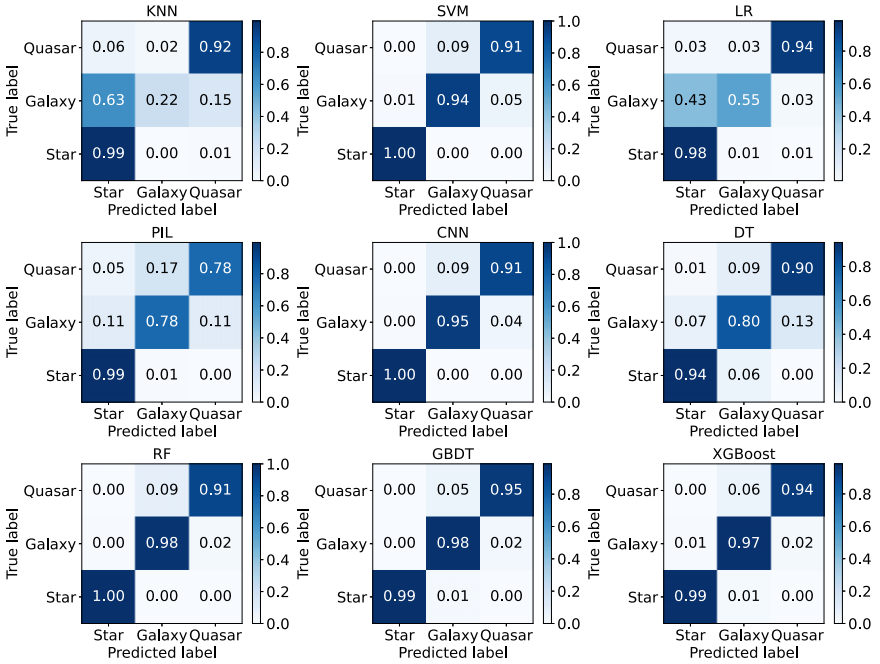

Figure 38. Accuracy of algorithms on star/galaxy/quasar on rest wavelength frame spectra. X-axis represents predicted labels conducted by experiments. Y-axis represents true labels of spectra. Figures in the grids are the consistent probabilities between predicted labels and true labels. Algorithm names are presented on the above of each confusion matrix.

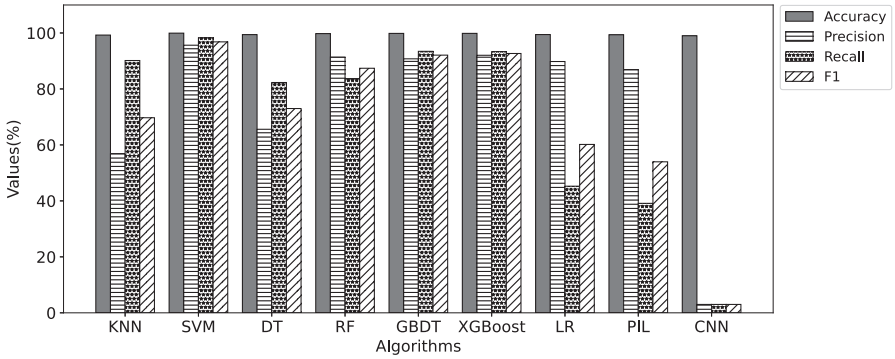

Figure 39. Results of algorithms on carbon stars. Four bars are four evaluation criteria of classification results.

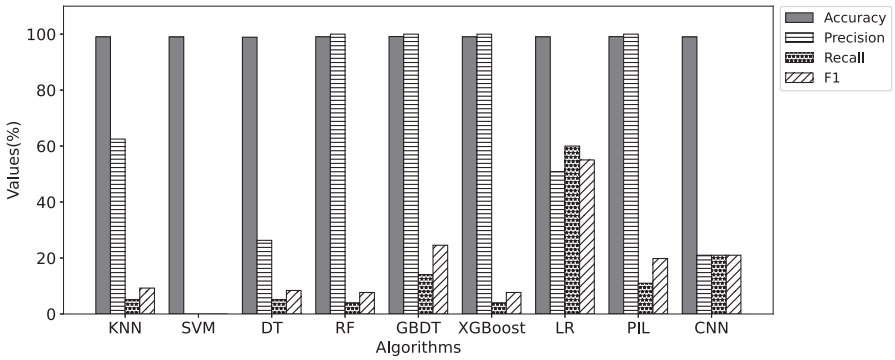

Figure 40. Results of algorithms on double stars. Four bars are four evaluation criteria of classification results.

It is difficult to identify carbon stars, double stars, and artefacts due to the unbalanced data distributions. Among these three rare objects, the performance of identifying carbon stars is better than others due to their obvious characteristics. The performance of searching for double stars is the worst. In short, researchers need to find other methods for rare object identification.

In this paper, we only evaluate the classification performance of nine basic algorithms on astronomical spectra. Other effective

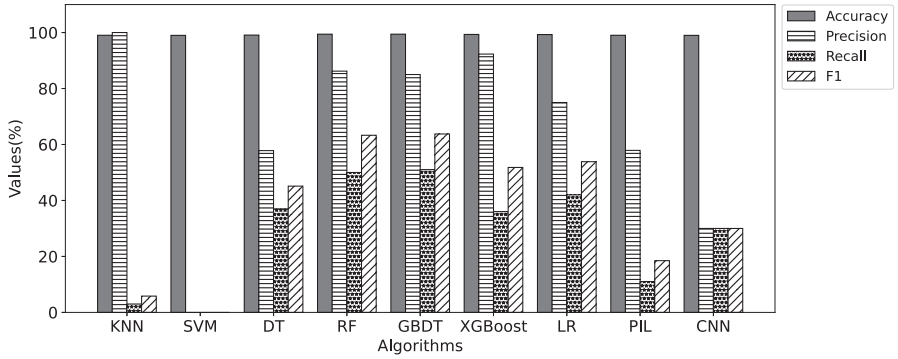

Figure 41. Results of algorithms on identifying artefacts. Four bars are four evaluation criteria of classification results.

Table 9. Source codes notes of classification algorithms.

Algorithms	KNN	SVM	LR	DT	RF	GBDT	XGBoost	CNN	PIL
Source files	KNN.py	SVM.py	LR.py	DT.py	RF.py	GBDT.py	XGBoost.py	CNN.py	PIL.py
Python version	python3.8								
Dependent packages	NUMPY; PANDAS; SKLEARN; SCIPY; PYTORCH								

methods still need to be analysed in the future. And experimental results in this paper can only provide a reference to researchers. In practical application scenarios, researchers need to choose appropriate methods according to their data characteristics.

ACKNOWLEDGEMENTS

The authors wish to thank the reviewer, Igor V Chilingarian, for his very helpful comments and suggestions.

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

The work is supported by the National Natural Science Foundation of China (Grant No. U1931209), Key Research and Development Projects of Shanxi Province (Grant No. 201903D121116), and the central government guides local Science and Technology Development Funds (Grant No. 20201070). Fundamental Research Program of Shanxi Province (Grant Nos. 20210302123223, 202103021224275).

DATA AVAILABILITY

Experimental data used for this work is obtained from The Guo Shou Jing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST) Data Release 8 (http://www.lamost.org/lmusers/) and Sloan Digital Sky Survey (SDSS) Data Release 16 (https://www.sdss.org/). Codes used in this paper is also available online at https://github.com/shichenhui/SpectraClassification.

REFERENCES

Aghanim N. et al., 2015, A&A, 580, A138
Agnello A., 2017, MNRAS, 471, 2013
Akras S., Leal-Ferreira M. L., Guzman-Ramirez L., Ramos-Larios G., 2019, MNRAS, 483, 5077
Almeida J. S., Aguerri J. A. L., Muñoz-Tuñón C., de Vicente A., 2010, ApJ, 714, 487
Arsioli B., Dedin P., 2020, MNRAS, 498, 1750
Astsatryan H., Gevorgyan G., Knyazyan A., Mickaelian A., Mikayelyan G. A., 2021, Astron. Comput., 34, 100442
Bae J.-M., 2014, Epidemiol. Health, 36, e2014025
Bai Y., Liu J., Wang S., Yang F., 2019, AJ, 157, 9
Baqui P. O. et al., 2021, A&A, 645, A87
Baran A. S., Sahoo S. K., Sanjayan S., Ostrowski J., 2021, MNRAS, 503, 3828
Baron D., 2019, preprint (arXiv:1904.07248)
Barrientos A., Solar M., Mendoza M., 2020, in Ballester P., Ibsen J., Solar M., Shortridge K., eds, ASP Conf. Ser.Vol. 522, Astronomical Data Analysis Software and Systems XXVII. Astron. Soc. Pac., San Francisco, p. 385
Bergen K. J., Johnson P. A., Maarten V., Beroza G. C., 2019, Science, 363, eaau0323
Biau G., Scornet E., 2016, Test, 25, 197
Bolton A. S. et al., 2012, AJ, 144, 144
Borne K. D., Vedachalam A., 2012, in Eric D., Feigelson G. J. B., eds, Statistical Challenges in Modern Astronomy V. Springer, Center for Astrostatistics, Penn State University, New York, p. 275
Brice M., Andonie R., 2019a, in Wang D., Doya K., eds, 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, Budapest, Hungary, p. 1
Brice M. J., Andonie R., 2019b, AJ, 158, 188
Bu Y., Zeng J., Lei Z., Yi Z., 2019, ApJ, 886, 128
Cabayol L. et al., 2019, MNRAS, 483, 529
Cai J., Yang Y., Yang H., Zhao X., Hao J., 2022, ACM Trans. Knowl. Discov. Data, 16, 1
Chao L., Wen-hui Z., Ji-ming L., 2019, Chin. Astron. Astrophys., 43, 539
Chao L., Wen-hui Z., Ran L., Jun-yi W., Ji-ming L., 2020, Chin. Astron. Astrophys., 44, 345
Chen Y. C., 2021, ApJS, 256, 34
Chen T., Guestrin C., 2016, in Krishnapuram B., Shah M., eds, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco, California, p. 785
Chi H., Li Z., Zhao W., 2022, in Li X., ed., Advances in Intelligent Automation and Soft Computing. Springer International Publishing, Cham, p. 495
Clarke A. O., Scaife A. M. M., Greenhalgh R., Griguta V., 2020, A&A, 639, A84
Cotar K. et al., 2019, MNRAS, 483, 3196
Czajkowski M., Grześ M., Kretowski M., 2014, Artif. Intell. Med., 61, 35

Daniel S. F., Connolly A., Schneider J., Vanderplas J., Xiong L., 2011, AJ, 142, 203
Davison W., Parkinson D., Tucker B. E., 2022, ApJ, 925, 186
Deng Z., Zhu X., Cheng D., Zong M., Zhang S., 2016, Neurocomputing, 195, 143
Dong H., Pan J., 2020, J. Phys.: Conf. Ser., 1624, 032017
Du B., Luo A., Zhang J., Wu Y., Wang F., 2012, in Radziwill N. M., Chiozzi G., eds, Proc. SPIE Conf. Ser. Vol. 8451, Software and Cyberinfrastructure for Astronomy II. SPIE, Bellingham, p. 845137
Du C., Luo A., Yang H., Hou W., Guo Y., 2016, PASP, 128, 034502
Duan F.-Q., Liu R., Guo P., Zhou M.-Q., Wu F.-C., 2009, Res. Astron. Astrophys., 9, 341
Farr J., Font-Ribera A., Pontzen A., 2020, J. Cosmol. Astropart. Phys., 2020, 015
Flores R. M., Corral L. J., Fierro-Santillán C. R., 2021, preprint (arXiv:2105.07110)
Franco-Arcega A., Flores-Flores L., Gabbasov R. F., 2013, in Félix Castro A. G., Mendoza M. G., eds, 2013 12th Mexican International Conference on Artificial Intelligence. IEEE Computer Society, Mexico City, Mexico, p. 181
Fremling C. et al., 2021, ApJ, 917, L2
Freund Y., Mason L., 1999, in Ivan Bratko S. D., ed., Proceedings of the Sixteenth International Conference on Machine Learning. ICML '99. Morgan Kaufmann Publishers Inc., San Francisco, CA, p. 124
Friedman J. H., 2001, Ann. Stat., 29, 1189
Fuqiang C., Yan W., Yude B., Guodong Z., 2014, Publ. Astron. Soc. Aust., 31, e001
Fushiki T., 2011, Stat. Comput., 21, 137
Gao Q., Shi J.-R., Yan H.-L., Yan T.-S., Xiang M.-S., Zhou Y.-T., Li C.-Q., Zhao G., 2019, ApJS, 245, 33
Gigoyan K. S., Russeil D., Mickaelian A. M., Sarkissian A., Avtandilyan M. G., 2012, A&A, 544, A95
Govada A., Gauri B., Sahay S., 2015, in Mauri J. L., Thampi S. M., Wozniak M., Marques O., eds, 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, Kochi, India, p. 258
Gray R. O., Corbally C. J., 2014, AJ, 147, 80
Green P., 2013, ApJ, 765, 12
Guo Z., Martini P., 2019, ApJ, 879, 72
Guo Y. X. et al., 2019, MNRAS, 485, 2167
Guo J., Zhao J., Zhang H., Zhang J., Bai Y., Walters N., Yang Y., Liu J., 2022, MNRAS, 509, 2674
Guzmán A. P. A., Esquivel A. E. O., Hernández R. D., Robles L. A., 2018, in Batyrshin I., ed., 2018 Seventeenth Mexican International Conference on Artificial Intelligence (MICAI). IEEE, Guadalajara, Mexico, p. 81
Harris C. R. et al., 2020, Nature, 585, 357
Hosenie Z., Lyon R., Stappers B., Mootoovaloo A., McBride V., 2020, MNRAS, 493, 6050
Hou W., Luo A.-l., Li Y.-B., Qin L., 2020, AJ, 159, 43
Hoyle B., Rau M. M., Paech K., Bonnett C., Seitz S., Weller J., 2015, MNRAS, 452, 4183
Hu Z., Chen J., Jiang B., Wang W., 2021, Universe, 7, 438
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ivanov S., Tsizh M., Ullmann D., Panos B., Voloshynovskiy S., 2021, Astron. Comput., 36, 100473
Jiang B., Wei D., Liu J., Wang S., Cheng L., Wang Z., Qu M., 2020, Universe, 6, 60
Jiang B., Fang X., Liu Y., Wang X., Liu J., 2021, Adv. Astron., 2021, 6748261
Jing-Min Z., Chen-Ye M., Lu W., Li-Ting D., Ting-Ting X., Lin-Pin A., Wei-Hong Z., 2020, Chin. Astron. Astrophys., 44, 334
Jingyi Y., Li Z., Chengjin Z., Jiaqi L., Zengjun B., 2018, in Liu Li Z. Y., ed., 2018 IEEE International Conference on Information and Automation (ICIA). IEEE, Wuyishan, China, p. 1290
Juvela M., 2016, A&A, 593, A58
Kang X., He S.-Y., Zhang Y.-X., 2021, Res. Astron. Astrophys., 21, 169
Karpov S. V., Malkov O. Y., Zhao G., 2021, MNRAS, 505, 207
Kerby S. et al., 2021, ApJ, 923, 75
Kesseli A. Y., West A. A., Veyette M., Harrison B., Feldman D., Bochanski J. J., 2017, ApJS, 230, 16
Khorrami Z. et al., 2021, A&A, 649, L8
Kong X., Luo A.-L., Li X.-R., Wang Y.-F., Li Y.-B., Zhao J.-K., 2018, PASP, 130, 084203
Kou S., Chen X., Liu X., 2020, ApJ, 890, 177
Kyritsis E., Maravelias G., Zezas A., Bonfini P., Kovlakas K., Reig P., 2022, A&A, 657, A62
Li J. et al., 2016, Res. Astron. Astrophys., 16, 110
Li X.-B., 2005, Decis. Support Syst., 41, 112
Li Y.-B. et al., 2018, ApJS, 234, 31
Li X.-R., Lin Y.-T., Qiu K.-B., 2019, Res. Astron. Astrophys., 19, 111
Liu Z., 2021, New Astron., 88, 1613
Liu C. et al., 2015b, Res. Astron. Astrophys., 15, 1137
Liu Z., Song L., Zhao W., 2016, MNRAS, 455, 4289
Liu W. et al., 2019, MNRAS, 483, 4774
Liu Z.-b., Zhao W.-j., 2017, Ap&SS, 362, 98
Liu Z.-b., Zhou F.-x., Qin Z.-t., Luo X.-g., Zhang J., 2018, Astrophys. Space Sci., 363, 140
Liu X.-W., Zhao G., Hou J.-L., 2015a, Res. Astron. Astrophys., 15, 1089
Lloyd Evans T., 2010, J. Astrophys. Astron., 31, 177
Lu Y., Pan J., Yi Z., 2020, in Long J., Pu Z., eds, 2020 Prognostics and Health Management Conference (PHM-Besançon). IEEE, Besançon, France, p. 366
Luo A.-L., Zhang Y.-X., Zhao Y.-H., 2004, in Lewis H., Raffi G., eds, Proc. SPIE Conf. Ser. Vol. 5496, Advanced Software, Control, and Communication Systems for Astronomy. SPIE, Bellingham, p. 756
Luo A.-L., Wu Y., Zhao J., Zhao G., 2008, in Bridger A., Radziwill N. M., eds, Proc. SPIE Conf. Ser. Vol. 7019, Advanced Software and Control for Astronomy II. SPIE, Bellingham, p. 1055
Luo A. et al., 2013, Proc. IAU Symp. 298, Setting the scene for Gaia and LAMOST. Cambridge Univ. Press, Cambridge, p. 428
Luo A. L. et al., 2015, Res. Astron. Astrophys., 15, 1095
Lupton R. H., Ivezic Z., Gunn J. E., Knapp G., Strauss M. A., Yasuda N., 2002, in Tyson J. A., Wolff S., eds, Proc. SPIE Conf. Ser. Vol. 4836, Survey and Other Telescope Technologies and Discoveries. SPIE, Bellingham, p. 350
Małek K. et al., 2013, A&A, 557, A16
Maravelias G., Bonanos A. Z., Tramper F., de Wit S., Yang M., Bonfini P., 2022, A&A, 666, A122
Margalef-Bentabol B., Huertas-Company M., Charnock T., Margalef-Bentabol C., Bernardi M., Dubois Y., Storey-Fisher K., Zanisi L., 2020, MNRAS, 496, 2346
Martins F., 2018, A&A, 616, A135
Maschmann D., Melchior A.-L., Mamon G. A., Chilingarian I. V., Katkov I. Y., 2020, A&A, 641, A171
Masters D., Capak P., 2011, PASP, 123, 638
Moraes R., Valiati J. F., Gavião Neto W. P., 2013, Expert Syst. Appl., 40, 621
Morice-Atkinson X., Hoyle B., Bacon D., 2018, MNRAS, 481, 4194
Muthukrishna D., Parkinson D., Tucker B. E., 2019, ApJ, 885, 85
Pattnaik R., Sharma K., Alabarta K., Altamirano D., Chakraborty M., Kembhavi A., Méndez M., Orwat-Kapola J. K., 2021, MNRAS, 501, 3457
Peng N., Zhang Y., Zhao Y., 2013, Sci. China Phys. Mech. Astron., 56, 1227
Pérez-Galarce F., Pichara K., Huijse P., Catelan M., Mery D., 2021, MNRAS, 503, 484
Pérez-Ortiz M. F., García-Varela A., Quiroz A. J., Sabogal B. E., Hernández J., 2017, A&A, 605, A123
Pichara K., Protopapas P., León D., 2016, ApJ, 819, 18
Podorvanyuk N., Chilingarian I., Katkov I., 2016, in Lorente N. P. F., Shortridge K., Wayth R., eds, ASP Conf. Ser. Vol. 512, Astronomical Data Analysis Software and Systems XXV. Astron. Soc. Pac., San Francisco, p. 253
Portillo S. K. N., Parejko J. K., Vergara J. R., Connolly A. J., 2020, AJ, 160, 45
Pruzhinskaya M. V., Malanchev K. L., Kornilov M. V., Ishida E. E. O., Mondon F., Volnova A. A., Korolev V. S., 2019, MNRAS, 489, 3591
Quinlan J. R., 1996, ACM Comput. Survey, 28, 71

Qu C.-X., Yang H.-F., Cai J.-H., Xun Y., 2020, Spectrosc. Spectral Anal., 40, 1304
Ramírez-Preciado V. G., Roman-Lopes A., Román-Zúñiga C. G., Hernández J., García-Hernández D., Stassun K., Stringfellow G. S., Kim J. S., 2020, ApJ, 894, 5
Rastegarnia F., Mirtorabi M. T., Moradi R., Sadr A. V., Wang Y., 2022, MNRAS, 511, 4490
Reis I., Poznanski D., Baron D., Zasowski G., Shahaf S., 2018, MNRAS, 476, 2117
Rosenfeld A., Vanderbrug G., 1977, IEEE Trans. Comput., 26, 384
Saez C. et al., 2015, MNRAS, 450, 2615
Sako M. et al., 2018, PASP, 130, 064002
Sharma K., Kembhavi A., Kembhavi A., Sivarani T., Abraham S., Vaghmare K., 2020, MNRAS, 491, 2280
Shi F., Liu Y.-Y., Kong X., Chen Y., 2014, A&A, 562, A36
Si J.-M. et al., 2015, Res. Astron. Astrophys., 15, 1671
Škoda P., Podsztavek O., Tvrdík P., 2020, A&A, 643, A122
Solarz A. et al., 2012, A&A, 541, A50
Solarz A., Bilicki M., Gromadzki M., Pollo A., Durkalec A., Wypych M., 2017, A&A, 606, A39
Solarz A. et al., 2020, A&A, 642, A103
Song W., Wang H., Maguire P., Nibouche O., 2018, Chemometr. Intell. Lab. Syst., 182, 79
Sookmee P., Suwannajak C., Techa-Angkoon P., Panyangam B., Tanakul N., 2020, in Lursinsap C., ed., 17th International Joint Conference on Computer Science and Software Engineering (JCSSE'20). IEEE, Bangkok, Thailand, p. 98
SubbaRao M., Frieman J., Bernardi M., Loveday J., Nichol B., Castander F., Meiksin A., 2002, in Starck J.-L., Murtagh F. D., eds, Proc. SPIE Conf. Ser. Vol. 4847, Astronomical Data Analysis II. SPIE, Bellingham, p. 452
Syarif I., Prügel-Bennett A., Wills G. B., 2016, TELKOMNIKA Telecommun. Comput. Electron. Control, 14, 1502
Tan L., Mei Y., Liu Z., Luo Y., Deng H., Wang F., Deng L., Liu C., 2022, ApJS, 259, 5
Tao Y., Zhang Y., Cui C., Zhang G., 2018, preprint (arXiv:1801.04839)
Tsalmantza P. et al., 2012, A&A, 537, A42
Vasconcellos E. C., de Carvalho R. R., Gal R. R., LaBarbera F. L., Capelato H. V., Frago Campos Velho H., Trevisan M., Ruiz R. S. R., 2011, AJ, 141, 189
Vilavicencio-Arcadia E., Navarro S. G., Corral L. J., Martínez C. A., Nigoche A., Kemp S. N., Ramos-Larios G., 2020, Math. Probl. Eng., 2020, 1751932
Wallerstein G., Knapp G. R., 1998, ARA&A, 36, 369
Wang K., Guo P., Luo A. L., 2017, MNRAS, 465, 4311
Wang L.-L., 2019, PASP, 131, 077001
Wang L.-L. et al., 2018, MNRAS, 474, 1873
Wei P. et al., 2014, AJ, 147, 101
Westfall K. B. et al., 2019, AJ, 158, 231
Xiao-Qing W., Jin-Meng Y., 2021, Chin. J. Phys., 69, 303
Yang Y., Cai J., Yang H., Zhang J., Zhao X., 2020, Expert Syst. Appl., 139, 112846
Yang P., Yang G., Zhang F., Jiang B., Wang M., 2021, Arch. Comput. Methods Eng., 28, 917
Yang H., Shi C., Cai J., Zhou L., Yang Y., Zhao X., He Y., Hao J., 2022a, MNRAS, 517, 5496
Yang Y., Cai J., Yang H., Li Y., Zhao X., 2022b, Expert Syst. Appl., 201, 117018
Yang Y., Cai J., Yang H., Zhao X., 2022c, Inf. Sci., 596, 414
Yi Z. et al., 2014, AJ, 147, 33
Yude B., Jingchang P., Bin J., Fuqiang C., Peng W., 2013, Publ. Astron. Soc. Aust., 30, e24
Yue L., Yi Z., Pan J., Li X., Li J., 2021, Optik, 225, 165535
Zhang M.-L., Zhou Z.-H., 2007, Pattern Recogn., 40, 2038
Zhang L. et al., 2016, New Astron., 44, 66
Zhang Y., Zhao Y., Wu X.-B., 2021, MNRAS, 503, 5263
Zhang B. et al., 2022, ApJS, 258, 26
Zhao Y., Zhang Y., 2008, Adv. Space Res., 41, 1955
Zhao G., Zhao Y.-H., Chu Y.-Q., Jing Y.-P., Deng L.-C., 2012, Res. Astron. Astrophys., 12, 723
Zhao Z., Wei J., Jiang B., 2022, Adv. Astron., 2022, 4489359
Zheng Z., Qiu B., 2020, J. Phys.: Conf. Ser., 1626, 012017
Zheng Z.-P., Qiu B., Luo A.-L., Li Y.-B., 2020, PASP, 132, 024504
Zhong J. et al., 2015a, Res. Astron. Astrophys., 15, 1154
Zhong J. et al., 2015b, AJ, 150, 42
Zou Z., el al., 2020, PASP, 132, 044503
Zou Z., Zhu T., Xu L., 2019, in Wang Y., Song W. W., eds, 4th International Conference on Computational Intelligence and Applications (ICCIA). IEEE, Nanchang, Jiangxi Province, p. 68

This paper has been typeset from a TEX/LATEX file prepared by the author.