Chinese Journal of Physics 69 (2021) 303–311

Contents lists available at ScienceDirect

Chinese Journal of Physics

journal homepage: www.sciencedirect.com/journal/chinese-journal-of-physics

Classification of star/galaxy/QSO and star spectral types from LAMOST data release 5 with machine learning approaches

Wen Xiao-Qing *, Yang Jin-Meng

Science of School, Nanchang University, Nanchang, 330031, PR China

ARTICLE INFO

Keywords:
stars
classification
random forest
LAMOST

ABSTRACT

We use 343,747 sources from LAMOST DR5 to do star/galaxy/QSO classification with machine learning approaches. Specifically, the 312,767 spectral labeled stars (G, K, M, F, A) are used to do star classification. The photometry of u, g, r, i, z, J, and H are used as machine learning features. For star/galaxy/QSO classification, the k nearest neighbor algorithm (KNN), decision tree (DT), random forest (RF) and support vector machine (SVM) perform well. For star classification, the accuracy of RF and SVM classification are higher than the accuracy of KNN and DT. The area under receiver operating characteristic curves of the four models are approaching to 1. The accuracy, precision, recall, f_score, Matthews correlation coefficient are always greater than 0.5. The four models perform all right in predicting the nature of sources and the star label.

1. Introduction

Some astronomical surveys face the challenge of classifying observed sources into stars, galaxies, QSOs or which spectral type the star is. It is becoming unfeasible for astronomers to manually verify and label individual sources. The manual labor is not expected to be able to keep up with the source counts anticipated for the next generation of telescopes. The development of machine learning algorithms has accelerated rapidly in the last decade, focusing on processing large data in high performance computing workflows and cloud computing systems [1,2].

Machine learning is already used in classifying galaxies and stars [3,4], verifying binaries [5], classing variable stars [6], searching groups or global stars [7,8], deriving stellar effective temperatures [9], verifying gravitational-wave [10], identifying strong lenses [11], classifying galaxy morphologies [12], and identifying carbon stars [13].

When spectroscopy is available, distinguishing astronomical source type is straightforward. However spectroscopy observation is time consuming and generally impractical for the largest samples of sources. In contrast, classifying sources using only photometry in multiple wavebands is comparatively fast. Photometry that can capture the overall shape of the spectrum to distinguish different types of sources has been demonstrated to be useful as machine learning features in source-type classification by a number of studies.

Costa-Duarte et al. [4] presented a star/galaxy classification for the Southern Photometric Local Universe Survey matched to Sloan Digital Sky Survey (SDSS)/DR13, based on a Machine Learning approach: the Random Forest algorithm. They indicated the broad photometric bands presented higher importance when compared to narrow ones. They used 12 bands from uJAVA to zSDSS (3574-783A).

Clarke et al. [3] used 1.2 million spectroscopically labeled sources from the SDSS catalog to train an optimized random forest

* Corresponding author.
E-mail address: xqwen@ncu.edu.cn (W. Xiao-Qing).

https://doi.org/10.1016/j.cjph.2020.03.008
Received 18 December 2019; Received in revised form 1 March 2020; Accepted 7 March 2020
Available online 13 March 2020

classifier using photometry from the SDSS and Wide field Infrared Survey Explorer (WISE). Using a test dataset of a further 1.2 million spectroscopically confirmed sources they determined that the random forest achieved f-scores of 0.991, 0.950, and 0.975 for galaxies, quasars and stars, respectively. Then, they applied their machine learning model to 111 million previously unlabelled sources from the SDSS photometric catalog without existing spectroscopic observations. The classification probabilities from the random forest as a measure of the likelihood of an individual classification were good, which were in agreement with the f-score derived from a nearest neighbor search around each source.

2. Data and preprocessing

2.1. data

LAMOST is a quasi-meridian reflecting Schmidt telescope of∼4 m effective aperture and a field of view of 5° in diameter [14-16]. LAMOST uses 4000 fibers to obtain spectra covering the entire optical wavelength range (≃3700–9000A), at a resolving power R ∼ 1800. The LAMOST Regular Survey consists of two components [17]: the LAMOST Extra-Galactic Survey of galaxies (LEGAS) that aims at studying the large scale structure of the universe, and the LAMOST Experiment for Galactic Understanding and Exploration (LEGUE) that aims at obtaining millions of stellar spectra in order to study the structure and evolution of the Milky Way [18]. LEGUE is sub-divided into three sub-surveys: the spheroid, the anti-center [19-21] and the disc surveys. The five-year Regular Survey finished in June 16th 2017 and the spectra have been released internally to the Chinese scientific community through the DR5 of LAMOST. The raw spectra are processed with the LAMOST two-dimensional (2D) pipeline [22–24], which includes dark and bias subtractions, cosmic ray removal, one-dimensional (1D) spectral extraction, merging sub-exposures, and finally, splicing the sub-spectra from the blue and red channels of the spectrographs, respectively. The LAMOST 1D pipeline is then carried out to perform spectral classification.

Our data is from the Data release 5 version 3 of the Large Sky Area Multi-Object Fiber Spectroscopic telescope(LAMOST). We use the general catalog, which includes the following spectral types of stars: A, F, G, K and M. The spectral types are corresponding to the classes we employ. The photometry are from SDSS u, g, r, i, z, combining the two Micron All Sky Survey (2MASS) J, H. The 7 bands magnitudes are as machine learning features. The raw data is 2.46 G. We constrain the SNR of u, g, r, i, z, J, H greater than 3. We delete the magnitudes of J or K lower than 10 or equal to 10, that are quite different from SDSS magnitudes and seem false. We choose observations with greater signal-to-noise ratios in the z band when LAMOST repeatedly observes one source. The data is 19.2 M after clean.

For star/galaxy/QSO classification, we use 'class' as a classification target. The data size is shown in Fig. 1a. For star classification, we use the '*subclass*' feature as a classification target. Since there are too many categories in the '*subclass*', we integrate the '*subclass*' as follows in Table 1.

For example, if the '*subclass*' value of a data is G1 or G3, we both classify the data as class G. Based on the above division, we classify the data into five categories, that is class G, class K, class M, class F, class A. Fig. 1b shows us the size of each class for star classification.

2.2. data balance

From Fig. 1b, we can find that class M has too little data to be put into the training set but class G and class F have a lot of data. So our data is unbalanced. Data imbalance can affect the establishment of the model. Oversampling and under-sampling strategies are one of the basic ways to solve the problem of data imbalance. Oversampling increases the number of samples of a minority class, while

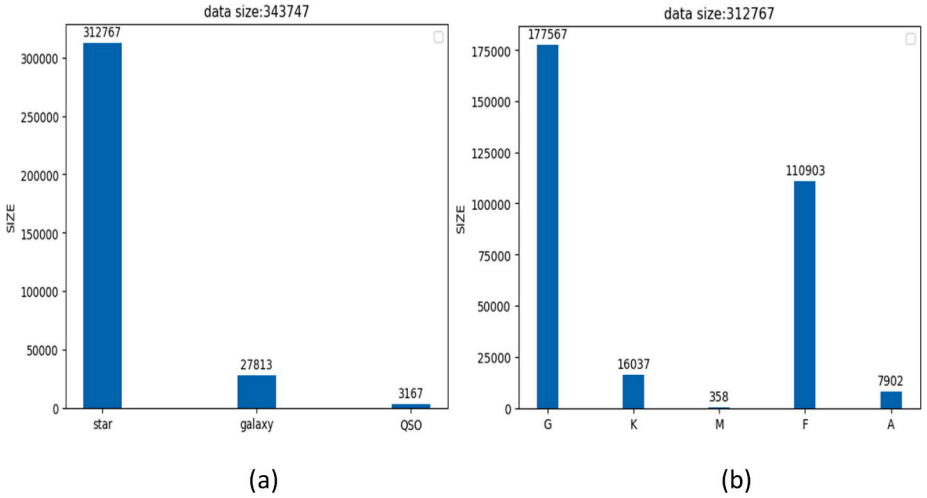

Fig. 1. (a) This histogram shows the number of star, galaxy and QSO; Fig. 1(b) This histogram shows the number of each class for star classification. The titles show the total amount of the samples.

Table 1
Data classification. The spectral types of A, F, G, K and M are corresponding to the classes.

Class G	Class K	Class M	Class F	Class A	
G3	K5	M0	F9	*A7V*	*A6V*
G7	K1	M2	F5	*A6IV*	*A7III*
G2	K3	M1	F7	*A1V*	*A8III*
G5	K7	M3	F2	*A2V*	*A3V*
G8	K0	M4	F6	*A5V*	*A0III*
G0	K4	M5	F0	*A7IV*	*A5*
G4	K2	M6	F8	*A9V*	*A0*
G9		M7	F3	*A2IV*	
G6		M8	F4	*A1IV*	
G1		M9		*A3IV*	

under-sampling reduces the majority of samples to obtain a relatively balanced data set. Synthetic Minority Oversampling Technique (SMOTE) [25] is often used to solve data imbalance problems. It is an improved scheme based on random oversampling algorithm, which adopts a strategy of simply copying samples to increase a few types of samples. However, it is easy to generate model over-fitting problems, that is, the information learned by the model is too special and not general enough. Still, the basic idea of the SMOTE algorithm is to analyze a small number of samples and manually synthesize new samples based on a few samples to add to the data set. In our work, we use the SMOTE algorithm to make the data balanced.

3. Machine learning algorithm and data reduction

3.1. Machine learning algorithm

In order to build our classification model, we use four supervised learning algorithms in machine learning which are k nearest neighbor algorithm (KNN), decision tree (DT), and random forest (RF), and kernel support vector machine (SVM).

- k nearest neighbor algorithm (KNN) [26]

KNN is arguably one of the simplest and most commonly used classification algorithms. In fact, when predicting a new value, the KNN is to predict the class of a new dot according to the class of k dots which are the nearest dots around it. So an important hyper-parameter of KNN is the number of neighbors *k*.

-decision tree (DT)

DT is a simple and widely used model. It can be used for classification and regression problems. The main function of DT is to summarize a series of decision rules by asking questions from a table with eigenvalues and labels. And it uses a tree diagram to present these decision rules.

DT usually have three steps: feature selection, decision tree generation, and cutting. Feature selection is a very important step before building a decision tree. If the features are randomly selected, the learning efficiency of the established decision tree will be greatly reduced. Usually when people select features, they will consider two different indicators: information gain and information gain ratio. The feature selection uses information entropy [27] and Gini index [28]. The goal of feature selection is to move in the direction of reduction of entropy and Gini index.

DT is easy to over-fit, generally requiring pruning [29], reducing the size of the tree structure, and alleviating over-fitting. People generally process the maximum depth and maximum number of features of the tree.

-random forest (RF)

RF is an extension of Bagging [30,31]. RF is an algorithm that integrates multiple trees through the idea of integrated learning. Its basic unit is the decision tree, and its essence belongs to a branch of machine learning - ensemble learning. RF has an important parameter: the number of trees. Other parameters are the same as decision trees.

-support vector machine (SVM) [32]

SVM is mainly used for the promotion of multi-classification in the field of pattern recognition, belonging to a supervised learning algorithm. Support vector machine classification model is referred to SVC. The important parameters of the kernel support vector machine are the regularization parameter C, the choice of the kernel, and the kernel-related parameters. We mainly use linear kernel and Gaussian kernel.

3.2. Data reduction

We use scikit-learn package in python to do data reduction. k-fold cross-validation divides data X into k equal and mutually exclusive subsets. That is, $X = X_1 \cup X_2 \cup X_3 \cup X_4 \cup \ldots \cup X_k$,$X_i \cap X_j = \varnothing (i \neq j)$. Each subset X_i can be obtained by hierarchical sampling from X. At each time we use k-1 subsets to train, and the remaining subset to validate. We do it for k times, $X_1, X_2, \ldots X_k$ are respectively used as validation set in turn. As a result, we can get k groups of training set and validation set. The final result is the average accuracy of predicting star spectral types of k groups in this work. We generally use cross-validation to optimize the hyper-parameters of the model. If the average accuracy value is higher, it means that the given parameter set is more reasonable at this time. The advantage of cross-validation is the ability to rule out accidental luck in selecting data. In other words, if we are very fortunate to put difficult types data into the training set in the random division, the accuracy of classification mode is unrealistically high. On the other hand, if we are unlucky to put similar data into the training set but the validation set has difficult types data, the classification accuracy is unrealistically low. Also, we try to let data be trained as more as possible via more fold cross-validation. For example, when using 10-fold cross-validation(k=10), we can use 90% of the data to train the model once a time. In general, the more data is, the more accurately model predict the result.

We have reprocessed the data following the steps.

Step 1: We divide the original data into training data and test data by 3:1.

Step 2: Balance training data by SMOTE and randomly select 50,000 data from the balanced training data for modeling, each class has 10,000 samples.

Step 3: Use KNN, DT, RF, SVM algorithm mentioned in machine learning to create four classification models. We apply the grid search-cross-validation method to optimize the hyper-parameter sets of four model. The 'best' set should be the one which maximizes the cross-validated accuracy. The grid search method is an exhaustive search method for specifying parameter values.

Step 4: Finally, the test data is input into four models for prediction. That shows the predictive ability of the model for unknown data.

3.3. Multiple classification evaluation method

Indicator 1: Confusion matrix

The confusion matrix is a visualization tool. In machine learning, especially in statistical classification, the confusion matrix is also called the error matrix.

Each column of the confusion matrix represents a predicted category. The total number of each column represents the sum of data predicted to be this category. Each row represents the true label of the data, and the total number of each row represents the sum of data belonging to this label. We are able to define the confusion matrix as M where $M_{i,j}$ is the number of samples actually belonging to class i but predicted in class j. Our confusion matrix is mainly used to view and evaluate the classification performance of unknown test set. The confusion matrix mainly contains three pieces of information: precision, recall rate, f-score(f-score is also called f-measure, which is the harmonic average of precision and recall). We can define the precision, recall rate, f-score as:

$$P(k) = \frac{M_{k,k}}{\sum_i M_{i,k}}, R(k) = \frac{M_{k,k}}{\sum_i M_{k,i}}, f - score(k) = 2 \cdot \frac{P(k) \cdot R(k)}{P(k) + R(k)} \tag{1}$$

The precision is based on our forecast results. It shows in the predicted positive category how many samples are actually positive category. The recall rate shows in the labeled category how many samples are predicted correctly. The precision and recall rate are a contradiction measure. Generally speaking, when the precision is high, the recall rate tends to be low; when the recall rate is high, the precision is often low. So we introduced comprehensive evaluation indicator: f-score. The closer three above values are to 1, the better the model performs.

Indicator 2: Accuracy (ACC)

Accuracy is the proportion of the correctly predicted samples to the total samples. Accuracy reflects the ability of the classifier to determine the entire sample. Accuracy is also based on the confusion matrix. Define the accuracy as

$$ACC = \frac{\sum_i M_{i,i}}{\sum_i \sum_j M_{i,j}} \tag{2}$$

The higher the accuracy is, the better the model performs on the data.

Indicator 3: Matthews correlation coefficient (MCC)

The Matthews correlation coefficient is the geometric mean of the problem and its dual regression coefficients. Single-class prediction is scored using the MCC for each class k, which is defined as:

$$MCC(k) = \frac{M_{k,k}n_k - o_k u_k}{\sqrt{\left(M_{k,k} + o_k\right)\left(M_{k,k} + u_k\right)(n_k + o_k)(n_k + u_k)}} \tag{3}$$

where $o_k = \sum_{i \neq k} M_{i,k}$, $u_k = \sum_{i \neq k} M_{k,i}$, $n_k = \sum_{i \neq k} \sum_{j \neq k} M_{i,j}$. The MCC returns a value between −1 and +1. A value of +1 indicates a perfect prediction, 0 indicates no better than a random prediction, and −1 indicates a complete inconsistency between prediction and observation. The closer the value is to 1, the more perfect the model is.

Indicator 4: Area under receiver operating characteristic curve(AUC) [33]

AUC is the estimation of whether the classifier works well. AUC stands for the incidence–the possibility that a sample is predicted correctly is greater than the possibility that the sample is predicted mistakenly. When AUC is 1, we have the perfect classifier. When AUC belongs to [0.85, 0.95], the classifier works well. When AUC is in a range of [0.7,0.85], the classifier is OK. When AUC is less than 0.7, the classifier has low effect or no effect.

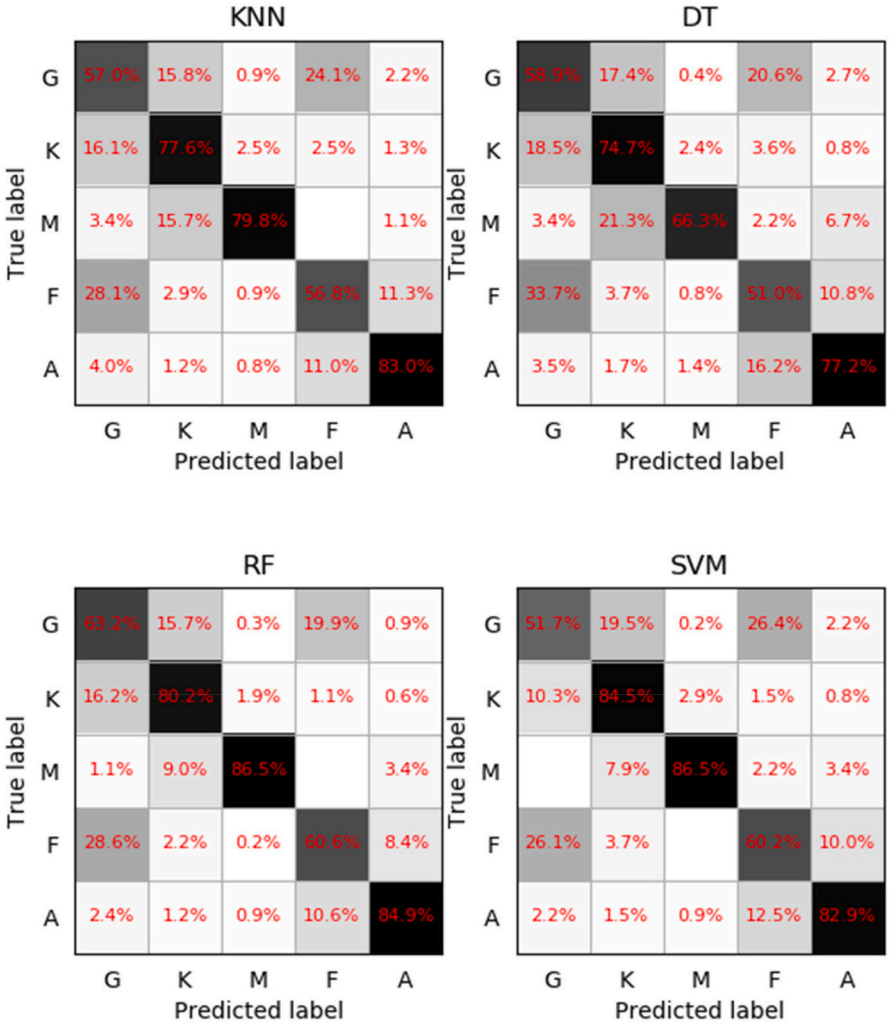

Fig. 2. This graphic is composed of four confusion matrix diagrams. A confusion matrix diagram show a classification model, the abscissa is the predicted label and the ordinate is the true label. The title of each confusion matrix shows the model name. The diagonal of the matrix shows the recall rate of each class.

4. Results

4.1. Star/galaxy/QSO classification on unknown test sets

We train the classifier using the training data and use the classifier to handle the test data. The 4 machine learning algorithm all perform very well on the classification of the source nature. The ACC is extremely high and greater than 95% for any of the four classifiers. Among them, RF performs the best, KNN performs the second. According to the recall rate, for the random forest model, 98.9% of the originally labeled star class data is predicted correctly. 97.8% of the originally labeled galaxy class is predicted correctly. 86.8% of the originally labeled QSO is predicted correctly. We can use the machine learning methods to classify the nature of sources practically.

4.2. Star spectral types classification on unknown test sets

After differentiating star from galaxy and QSO, we try to do the star spectral types classification. When we input the test set into the model for star spectral types, we can observe the classification of the test set by confusion matrix, as shown in Fig. 2.

For example, for the random forest model, the row 'True label=G' can be explained as follows: 63.2% of the originally labeled G class data is predicted correctly. However 15.7% of the originally labeled G class data is predicted to be K class, 0.3% is predicted to be M class, 19.9% is predicted to be F class, 0.9% is predicted to be A class. The column 'Predicted label=G' can be explained as follows: in the predicted G class stars, 63.2% are true G stars, 16.2%, 1.1%, 28.6%, 2.4% are actually labeled K, M, F, A stars, respectively. The diagonal numbers represent the recall rate, 63.2%, 80.2%, 86.5%, 60.6%, 84.9% for G, K, M, F, A stars, respectively. The average prediction rate of 5 type stars is 0.725.

There are similar explanations in the diagrams for KNN, DT, and SVC models. The RC and SVC model have better predictions for G, K, M, F, A type stars.

It can also be seen from matrix figure that the data of class G and class F are seriously cross misclassified. In the RF model, 19.9% of class G is predicted to be class F, while 28.6% of class F is predicted to be class G. It can also be seen from Table 2 that the prediction effects of class G and class F are not as good as the other three classes. We check the originally labeled G and F data, and find the 7 machine learning features are similar of the two classes. The mean values of the 7 features are 14.33, 13.54, 13.43, 13.94, 13.42, 12.47, 12.17 for G class. The standard deviation is 1.08, 0.98, 0.92, 1.12, 1.04, 1.22, 1.11, respectively. The mean values of the 7 features are 14.12, 13.45, 13.36, 13.69, 13.30, 12.87, 12.44 for F class. The standard deviation is 1.15, 1.04, 1.00, 1.07, 1.00, 2.08, 1.31, respectively. The features are similar for machine to distinguish.

We can also observe the detailed classification information of the four models in Table 3. Table 3 shows the ACC, precision, recall, f-

Table 2
results of four classifier for star/galaxy/QSO classification.

KNN ACC:0.970 **class**	**precision**	**Recall**	**f_score**	**MCC**	**AUC**
star	0.983	0.989	0.986	0.979	0.993
galaxy	0.963	0.978	0.97	0.957	0.985
QSO	0.946	0.844	0.892	0.885	0.949
avg/total	0.97	0.971	0.97	0.94	0.988
DT **ACC:0.958** **class**	**precision**	**Recall**	**f_score**	**MCC**	**AUC**
star	0.978	0.978	0.978	0.968	0.976
galaxy	0.957	0.959	0.958	0.939	0.958
QSO	0.87	0.856	0.863	0.851	0.913
avg/total	0.958	0.959	0.958	0.919	0.967
RF **ACC:0.973** **class**	**precision**	**Recall**	**f_score**	**MCC**	**AUC**
star	0.983	0.989	0.986	0.979	0.997
galaxy	0.965	0.978	0.971	0.959	0.991
QSO	0.96	0.868	0.912	0.906	0.973
avg/total	0.973	0.973	0.972	0.948	0.994
SVM **ACC:0.953** **class**	**precision**	**Recall**	**f_score**	**MCC**	**AUC**
star	0.988	0.964	0.976	0.965	0.995
galaxy	0.932	0.97	0.951	0.929	0.982
QSO	0.891	0.816	0.852	0.84	0.935
avg/total	0.954	0.953	0.953	0.911	0.985

Table. 3
Star classification report of four classification models.

KNN ACC:0.668 class	precision	recall	f_score	MCC	AUC
G	0.540	0.570	0.555	0.454	0.778
K	0.785	0.776	0.780	0.730	0.902
M	0.582	0.798	0.673	0.674	0.905
F	0.602	0.568	0.584	0.495	0.793
A	0.848	0.830	0.839	0.801	0.926
avg/total	0.691	0.688	0.689	0.631	0.873
DT ACC:0.655 class	**precision**	**recall**	**f_score**	**MCC**	**AUC**
G	0.513	0.589	0.548	0.444	0.756
K	0.752	0.747	0.749	0.692	0.859
M	0.541	0.663	0.596	0.590	0.830
F	0.557	0.510	0.532	0.434	0.773
A	0.838	0.772	0.804	0.761	0.877
avg/total	0.662	0.655	0.657	0.584	0.844
RF ACC:0.725 class	**precision**	**recall**	**f_score**	**MCC**	**AUC**
G	0.572	0.632	0.600	0.508	0.855
K	0.801	0.802	0.802	0.755	0.958
M	0.700	0.865	0.774	0.773	0.985
F	0.657	0.606	0.631	0.552	0.871
A	0.893	0.849	0.870	0.841	0.968
avg/total	0.730	0.725	0.727	0.686	0.940
SVM ACC:0.702 class	**precision**	**recall**	**f_score**	**MCC**	**AUC**
G	0.573	0.517	0.543	0.448	0.801
K	0.769	0.845	0.805	0.759	0.952
M	0.658	0.865	0.748	0.749	0.990
F	0.597	0.602	0.600	0.510	0.802
A	0.862	0.829	0.845	0.809	0.963
avg/total	0.699	0.702	0.699	0.655	0.911

score, MCC, and AUC of the four models. As can be seen from Fig. 2 and Table 3, the accuracy of the RF and SVM are higher than the accuracy of the KNN and DT. Among the four models, the RF classification model performs the best and the DT performs the worst. The RF algorithm cost quite less time than SVM. For A stars, the prediction rate is the highest. For K stars, the prediction rate is the second highest. There are enough sample of the two type stars to train the model. So we get the highest prediction rate. Though we balance the data, the information to train the model is not enough. SMOTE algorithm needs to synthesize new samples based on a few samples to add to the data set. New samples may be similar to the sample used for creation. So, M stars get the lower prediction rate than A and K stars for too little sample although SMOTE improve the situation. It is strange that G and F stars have the biggest sample but give the lower prediction rate. We find the models always mix the G and F stars by mistakes. We carefully check the data and find that the 7 features are similar for G and F stars. That may lead to the misclassification of G and F stars. Though SMOTE affects the prediction rate to a certain extent, the similar features are the main reason. That the AUC of the four models is approaching to 1 illustrates the four models performs all right in predicting star label. The ACC, precision, recall, f_score, and MCC are always greater than 0.5.

In addition, the interstellar extinction may have effect on the classification. The estimation of effective temperature and interstellar extinction are always hard nut to crack for recent surveys, such as GAIA, LAMOST and APOGEE. These two parameters always couple together. Using the extinction law in Indebetouw et al. [34] and the mean stellar colors from isochrones in Girardi et al. [35], one may estimate the extinction in the Ks band: $A(K_S) = 1.05(J - H - 0.76)$. In this paper, we use the u, g, r, i, z, J and H magnitudes to be machine learning features. The J and H magnitudes have already be included. If the extinction $A(K_S)$ have correlation with the J and H magnitudes, we can receive that the extinction information has already been included in the 7 features. Except for another correlation by Indebetouw et al. [34] and Girardi et al. [35] $A(K_S) = 1.82(H - K_S - 0.13)$, Majewski et al. [36] give the extinction by $A(K_S) = 0.918(H - [4.5\mu] - 0.08)$. Even though these correlations for extinction are not established, we also experiment to add K_S and $[4.5\mu]$ to be our machine learning features by cross-match with WISE. The classification results for star/galaxy/QSO and for star spectral types using 9 features are similar to that using 7 features. One possible explanation is that the LAMOST DR5 catalog only carries stars with negligible extinction. We will discuss more in Wen et al. 2020 (prepared) about the similar results when selecting different amount of features for effective temperature regression.

5. Conclusion

Machine learning algorithms can be used to predict the nature of sources and star spectral labels. The features of u, g, r, i, z, J, H magnitudes can be used to precisely predict star/galaxy/QSO classification. However in star classification, KNN, DT, RF, and SVM models can predict star spectral labels to some extent.

1 The ACC is extremely high and greater than 95% for any of the four classifiers in star/galaxy/QSO classification.
2 The accuracy of the RF and SVM are higher than the accuracy of the KNN and DT in star classification. Among the four models, RF is the best. The performance of DT is poorest.
3 The data of class G and class F are seriously cross misclassified so the prediction effects of the two types are not as good as those of the other three classes.
4 For A and K stars, we have the highest prediction rates. For M stars, the prediction rate is lower than those of A and K stars for few samples although SMOTE improve the situation.

Declaration of Competing Interest

Wen, Xiao-Qing designed the research. Yang, Jin-Meng performed the calculations. Wen, Xiao-Qing interpreted and analyzed the results. The authors declare no competing interests.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant nos. 62066026, 11563005) and Jiangxi Province Science Foundation for Youths(No. 20151BAB212011). This work has made use of data products from the Guoshoujing Telescope (the Large Sky Area Multi-Object Fibre Spectroscopic Telescope, LAMOST). 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 the National Astronomical Observatories, Chinese Academy of Sciences.

References

[1] N. Jones, Computer science: the learning machines, Nat. News 505 (2014) 146.
[2] C. Wu, R. Buyya, K. Ramamohanarao, Big data analytics= machine learning+ cloud computing, arXiv preprint arXiv:1601.03115, (2016).
[3] A.O. Clarke, A. Scaife, R. Greenhalgh, V. Griguta, Identifying galaxies, quasars and stars with machine learning: a new catalogue of classifications for 111 million SDSS sources without spectra,arXiv preprint arXiv:1909.10963, (2019).
[4] M.V. Costa-Duarte, L. Sampedro, A. Molino, H.S. Xavier, F.R. Herpich, A.L. Chies-Santos, C.E. Barbosa, A. Cortesi, W. Schoenell, A. Kanaan, The S-PLUS: a star/galaxy classification based on a Machine Learning approach, arXiv preprint arXiv:1909.08626, (2019).
[5] D. Chatterjee, S. Ghosh, P.R. Brady, S.J. Kapadia, A.L. Miller, S. Nissanke, F. Pannarale, A machine learning based source property inference for compact binary mergers, arXiv preprint arXiv:1911.00116, (2019).
[6] Z. Hosenie, R.J. Lyon, B.W. Stappers, A. Mootoovaloo, Comparing Multiclass, Binary, and Hierarchical Machine Learning Classification schemes for variable stars, Mon Not R Astron Soc 488 (2019) 4858–4872.
[7] W. Dobbels, M. Baes, S. Viaene, S. Bianchi, J.I. Davies, V. Casasola, C. Clark, J. Fritz, M. Galametz, F. Galliano, Predicting the global far-infrared SED of galaxies via machine learning techniques, arXiv preprint arXiv:1910.06330, (2019).
[8] J. Lee, I. Song, Evaluation of nearby young moving groups based on unsupervised machine learning, Mon Not R Astron Soc 489 (2019) 2189–2194.
[9] Y. Bai, J. Liu, S. Wang, F Yang, Machine Learning Applied to star-Galaxy-QSO classification and stellar effective temperature regression, Astron. J. 157 (2018) 9.
[10] R.E. Colgan, K.R. Corley, Y. Lau, I. Bartos, J.N. Wright, Z. Mi慢rka, S. Marka, Efficient gravitational-wave glitch identification from environmental data through machine learning, arXiv preprint arXiv:1911.11831, (2019).
[11] T. Cheng, N. Li, C.J. Conselice, A. Arag贸n-Salamanca, S. Dye, R.B. Metcalf, Identifying strong lenses with unsupervised machine learning using convolutional autoencoder, arXiv preprint arXiv:1911.04320, (2019).
[12] G. Martin, S. Kaviraj, A. Hocking, S.C. Read, J.E. Geach, Galaxy morphological classification in deep-wide surveys via unsupervised machine learning, Mon Not R Astron Soc 491 (2019) 1408–1426.
[13] Y. Li, A. Luo, C. Du, F. Zuo, M. Wang, G. Zhao, B. Jiang, H. Zhang, C. Liu, L. Qin, Carbon stars identified from LAMOST DR4 using machine learning, Astrophys. J. Suppl. Ser. 234 (2018) 31.
[14] X. Cui, Y. Zhao, Y. Chu, G. Li, Q. Li, L. Zhang, H. Su, Z. Yao, Y. Wang, X. Xing, The large sky area multi-object fiber spectroscopic telescope (LAMOST), Res Astron Astrophys 12 (2012) 1197.
[15] D. Su, X. Cui, Active optics in LAMOST, Chin. J. Astron. Astrophys. 4 (2004) 1.
[16] S. Wang, D. Su, Y. Chu, X. Cui, Y. Wang, Special configuration of a very large Schmidt telescope for extensive astronomical spectroscopic observation, Appl. Opt. 35 (1996) 5155–5161.
[17] G. Zhao, Y. Zhao, Y. Chu, Y. Jing, L. Deng, LAMOST spectral survey-An overview, Res Astron Astrophys 12 (2012) 723.
[18] L. Deng, H.J. Newberg, C. Liu, J.L. Carlin, T.C. Beers, L. Chen, Y. Chen, N. Christlieb, C.J. Grillmair, P. Guhathakurta, LAMOST Experiment for Galactic Understanding and Exploration (LEGUE)-The survey's science plan, Res Astron Astrophys 12 (2012) 735.
[19] X. Liu, H. Yuan, Z. Huo, L. Deng, J. Hou, Y. Zhao, G. Zhao, J. Shi, A. Luo, M. Xiang, LSS-GAC-A LAMOST spectroscopic survey of the galactic anti-center, Proc. Int. Astron. Union 9 (2013) 310–321.
[20] M. Xiang, X. Liu, H. Yuan, Z. Huo, Y. Huang, C. Wang, B. Chen, J. Ren, H. Zhang, Z. Tian, LAMOST Spectroscopic Survey of the Galactic Anticentre (LSS-GAC): the second release of value-added catalogues, Mon Not R Astron Soc 467 (2017) 1890–1914.
[21] H. Yuan, X. Liu, Z. Huo, M. Xiang, Y. Huang, B. Chen, H. Zhang, N. Sun, C. Wang, H. Zhang, LAMOST Spectroscopic Survey of the Galactic Anticentre (LSS-GAC): target selection and the first release of value-added catalogues, Mon Not R Astron Soc 448 (2015) 855–894.
[22] A. Luo, H. Zhang, Y. Zhao, G. Zhao, X. Cui, G. Li, Y. Chu, J. Shi, G. Wang, J. Zhang, Data release of the LAMOST pilot survey, Res Astron Astrophys 12 (2012) 1243.
[23] A. Luo, Y. Zhao, G. Zhao, L. Deng, X. Liu, Y. Jing, G. Wang, H. Zhang, J. Shi, X. Cui, The first data release (DR1) of the LAMOST regular survey, Res Astron Astrophys 15 (2015) 1095.
[24] Y. Song, A. Luo, G. Comte, Z. Bai, J. Zhang, W. Du, H. Zhang, J. Chen, F. Zuo, Y. Zhao, Relative flux calibration for the Guoshoujing Telescope (LAMOST), RES Astron Astrophys 12 (2012) 453.

[25] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J Artif Intell Res 16 (2002) 321–357.
[26] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer Sci. Bus. Media (2009).
[27] J.R. Quinlan, Induction of decision trees, Mach Learn 1 (1986) 81–106.
[28] L. Breiman, J. Friedman, R. Olshen, C Stone, Classification and Regression Trees, 37, Wadsworth Int, Group, 1984, pp. 237–251.
[29] J.R. Quinlan, Combining instance-based and model-based learning, 1993, pp. 236-243.
[30] L. Breiman, Bagging predictors, Mach Learn 24 (1996) 123–140.
[31] L. Breiman, Random forests, Mach Learn 45 (2001) 5–32.
[32] C. Hsu, C. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2002) 415–425.
[33] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Mach Learn 45 (2001) 171–186.
[34] R. Indebetouw, J.S. Mathis, B.L. Babler, M.R. Meade, C. Watson, B.A. Whitney, M.J. Wolff, M.G. Wolfire, M. Cohen, T.M. Bania, The wavelength dependence of interstellar extinction from 1.25 to 8.0 μm using GLIMPSE data, Astrophys. J. 619 (2005) 931.
[35] L. Girardi, G. Bertelli, A. Bressan, C. Chiosi, M. Groenewegen, P. Marigo, B. Salasnich, A. Weiss, Theoretical isochrones in several photometric systems-I. Johnson-Cousins-Glass, HST/WFPC2, HST/NICMOS, Washington, and ESO Imaging Survey filter sets, Astron Astrophys 391 (2002) 195–212.
[36] S.R. Majewski, G. Zasowski, D.L. Nidever, Lifting the dusty veil with near-and mid-infrared photometry. I. Description and applications of the Rayleigh-jeans color excess method, Astrophys. J. 739 (2011) 25.