Publications of the Astronomical Society of Australia (2023), **40**, e037, 10 pages
doi:10.1017/pasa.2023.35

CAMBRIDGE
UNIVERSITY PRESS

Research Article

A novel approach for variable star classification based on imbalanced learning

Jingyi Zhang[1], Yanxia Zhang[1], Zihan Kang[1], Changhua Li[2], Yihan Tao[2], Yongheng Zhao[1] and Xue-Bing Wu[3]

[1]Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China, [2]National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China and [3]Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing, China

Abstract

The advent of time-domain sky surveys has generated a vast amount of light variation data, enabling astronomers to investigate variable stars with large-scale samples. However, this also poses new opportunities and challenges for the time-domain research. In this paper, we focus on the classification of variable stars from the Catalina Surveys Data Release 2 and propose an imbalanced learning classifier based on Self-paced Ensemble (SPE) method. Compared with the work of Hosenie et al. (2020), our approach significantly enhances the classification *Recall* of Blazhko RR Lyrae stars from 12% to 85%, mixed-mode RR Lyrae variables from 29% to 64%, detached binaries from 68% to 97%, and LPV from 87% to 99%. SPE demonstrates a rather good performance on most of the variable classes except RRab, RRc, and contact and semi-detached binary. Moreover, the results suggest that SPE tends to target the minority classes of objects, while Random Forest is more effective in finding the majority classes. To balance the overall classification accuracy, we construct a Voting Classifier that combines the strengths of SPE and Random Forest. The results show that the Voting Classifier can achieve a balanced performance across all classes with minimal loss of accuracy. In summary, the SPE algorithm and Voting Classifier are superior to traditional machine learning methods and can be well applied to classify the periodic variable stars. This paper contributes to the current research on imbalanced learning in astronomy and can also be extended to the time-domain data of other larger sky survey projects (LSST, etc.).

Keywords: Astronomy databases: catalogue; method: statistical; method: machine learning; astronomical techniques: time-domain astronomy

(Received 26 September 2022; revised 2 June 2023; accepted 27 June 2023)

1. Introduction

Stellar luminosity varies slowly in the long-term evolution, but some stars exhibit noticeable luminosity changes on the short timescale (compared to the evolutionary timescale). These stars are called variable stars. Depending on the causes of variations, they are classified into intrinsic and extrinsic variables (Eyer & Mowlavi 2008). Among them, the periodic variables display regular or semi-regular luminosity changes, and their periods are usually correlated with the physical or geometric properties of variables. The mass, luminosity, radius, and age of variables can be inferred from their periods using empirical relations (Chen et al. 2018). Some variables, such as Cepheid stars and RR Lyrae stars, can serve as standard candles for distance measurement (Alloin & Gieren 2003). Moreover, as tracers, they reveal the structure, chemical, and dynamical evolution of the Galaxy, as well as the substructure of the halo (Zhang, Zhang, & Zhao 2018; Koposov et al. 2019; Price-Whelan et al. 2019; Prudil et al. 2020). Therefore, periodic variables are not only crucial for studying the structure of galaxies but also can be used to constrain the stellar evolution model.

The development of large-scale sky survey projects, such as the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010), Kepler (Koch et al. 2010), and Catalina Real-Time Transient Survey (CRTS; Drake et al. 2017), has enabled repeated observations of large areas of the sky every few nights. With the accumulation of abundant light curve data, the study of time-domain astronomy has become imperative. Time-domain astronomy is now entering a golden age, which spans across electromagnetic wavelengths from radio to gamma-rays (Graham et al. 2017). In this new era of data explosion, it is impractical to classify variable stars by visual inspection alone. Therefore, it is essential to achieve automatic classification of massive variable stars and assign the unprecedented large number of light curves to known or unknown classes. To better understand periodic variables, many previous works have focused on their classification. Chen et al. (2018) used the colours, periods, and shapes of WISE light curves to classify periodic variable stars by physical cuts. Petrosky et al. (2021) applied the non-parametric features of WISE light curves and physical cuts to distinguish periodic and aperiodic variable stars and discussed the classification of periodic variables. The increasing amount of data makes automatic classification necessary. At present, machine learning has been widely used to solve classification problems in astronomy, such as Support Vector Machine (SVM; Peng, Zhang, & Zhao 2013; Jin et al. 2019), Random Forest (Gao, Zhang, & Zhao 2009; Zhang, Zhao, & Wu 2021), and so on. However, traditional machine learning algorithms may not be

Corresponding author: Yanxia Zhang; Email: zyx@bao.ac.cn

Cite this article: Zhang J, Zhang Y, Kang Z, Li C, Tao Y, Zhao Y and Wu X-B. (2023) A novel approach for variable star classification based on imbalanced learning. *Publications of the Astronomical Society of Australia* **40**, e037, 1–10. https://doi.org/10.1017/pasa.2023.35

suitable for some cases, for example, when the dataset is extremely imbalanced. Standard machine learning algorithms assume that the number of samples belonging to different classes are roughly equal. Therefore, the uneven distribution of data can impair the performance of algorithms. The implicit optimisation goal behind the design of these learning algorithms is classification accuracy on the dataset, which causes the learning algorithm to be more biased towards the majority class.

There are three kinds of approaches for dealing with class imbalanced problems, namely data level, algorithm level, and ensemble methods (Liu et al. 2020). Data-level methods are the earliest and most widely used methods in the field of imbalanced learning, which are also called re-sampling methods. They aim to modify the training data to improve the performance of machine learning algorithms. Algorithm-level methods mainly adapt traditional machine learning algorithms to correct their preference for the majority class. The most popular branch of such methods is cost-sensitive learning. By incorporating the costs into the classifier construction, cost-sensitive learning makes the prediction more favourable to the minority (Zhang, Zhang, & Zhao 2020). Ensemble methods combine data-level or algorithm-level method with ensemble learning to obtain powerful ensemble classifiers. Ensemble learning is favoured due to its excellent performance on imbalanced tasks. Hoyle et al. (2015) used a tree-based data augmentation method for apparent magnitudes, which provided low-bias redshift estimation of galaxies. Hosenie et al. (2020, hereafter Ho20) classified the periodic variable stars of CRTS catalogs using imbalanced learning, including Synthetic Minority Oversampling Technique (SMOTE; Chawla et al. 2002) and hierarchical classifier. However, the algorithm can be further improved by addressing the sample imbalance and feature selection issues. Liu et al. (2020) proposed a new approach (Self-paced Ensemble; SPE) for imbalanced learning. Compared with the traditional methods, the SPE algorithm is an efficient, general-purpose, and robust ensemble imbalanced learning framework. In this paper, we will apply the SPE algorithm to classify the variables.

In this work, our main objective is to classify the periodic variables by imbalanced learning. The paper is organised as follows. Section 2 introduces the CRTS survey and describes how we obtain the samples. In Section 3, we review the advantages and disadvantages of different imbalanced learning approaches, explain the feature extraction of the sample and the hyperparameter optimisation of SPE, and define the evaluation metrics of classification performance. In Section 4, we apply the SPE approach, compare and discuss its performance with Ho20, and present a Voting Classifier, and then analyse the results. Finally, we conclude and outline future work.

2. The data

The CRTS (Drake et al. 2017) is an astronomical time-domain survey that covers 33000 deg^2 of the sky to discover rare and interesting transient phenomena. The survey used three dedicated telescopes of the highly successful Catalina Sky Survey (CSS) project to acquire data. Its limiting magnitude is about 20–21 mag per exposure with time baselines from 10 min to 6 yr. The survey has detected about 500 million sources, which is a scientific and technological test bed and precursor for the larger sky survey. It also has produced a catalogue of 11 classes of periodic variable stars for 6 yr of optical photometry. Here, we take these 11 classes into account from the Catalina Surveys Data Release 2[a] for our analysis, as shown in Table 1. For clarity, Fig. 1 presents folded light curves of different kinds of variable stars. As shown in Fig. 1, most kinds of variables have different light curve shapes and so they are easy to discriminate; while some kinds of variables have similar shapes, we need consider other parameters (e.g., period) if we want to separate them.

Table 1. The number of different classes of variables in the CRTS dataset.

Classes of variable stars	No.
RRab	4325
RRc	3752
RRd	502
Blazhko	171
Contact & Semi-Detached Binary	18803
Detached Binary	4509
Rotational	3636
Long Period Variable	1286
Delta-Scuti	147
Anomalous Cepheid	153
Type-II Cepheid	153

3. The method

3.1. Imbalanced learning

The existing traditional imbalanced learning algorithms can be categorised into three types: data-level, algorithm-level and integration-level methods. The advantages and disadvantages (Liu et al. 2020) of each method are summarised in Table 2. The main reason why the existing methods fail in such tasks is that they ignore the difficulties inherent in the nature of imbalanced learning. These difficulties may arise from the data collection process (such as noise and missing values), or from the characteristics of the dataset (such as class overlap and large data volume), or from the machine learning model (the model capacity is too small or too large) and the task itself (class imbalance). Besides the class imbalance, these factors also significantly degrade the classification performance. Their impact can be further amplified by the high imbalance ratio. Traditional imbalanced learning methods usually only address one or several of these factors, and the final performance depends on the choice of hyperparameters. For instance, the number of neighbours considered in the distance-based re-sampling will affect the sensitivity to noise (Liu et al. 2020). The cost matrix in the cost-sensitive learning needs to be set by experts. Since SPE (Liu et al. 2020) does not require any predefined distance metric or computation, it is more convenient for application and more efficient for computation. Moreover, SPE is adaptive to different models and robust to noises and missing values.

In this paper, we use SPE to classify the variables. Fig. 2 illustrates the SPE framework. Compared with traditional methods, it has some advantages as follows (Liu et al. 2020):

[a] Catalina Surveys Data Release 2.

https://doi.org/10.1017/pasa.2023.35 Published online by Cambridge University Press

Table 2. Comparison of different imbalance methods.

Comparison of different methods			
	Data level	Algorithm level	Integration level
Types	Under-sampling Oversampling Oversampling + under-sampling	Cost-sensitive learning	SMOTE-Boost SMOTE-Bagging Easy-Ensemble, etc.
Pros	It can denoise and balance class distribution	Increasing training complexity can be directly used for multi-classification problems	Its performance is generally good and can be dynamically adjusted using feedback during iterations
Cons	Sampling process is computationally inefficient; susceptible to noise; oversampling methods generate too much data; unsuitable for complex datasets where distance cannot be calculated	Requires prior domain knowledge; cannot generalise to different tasks; depends on specific classifiers	Oversampling and ensemble further increase computational overhead; not robust to noise

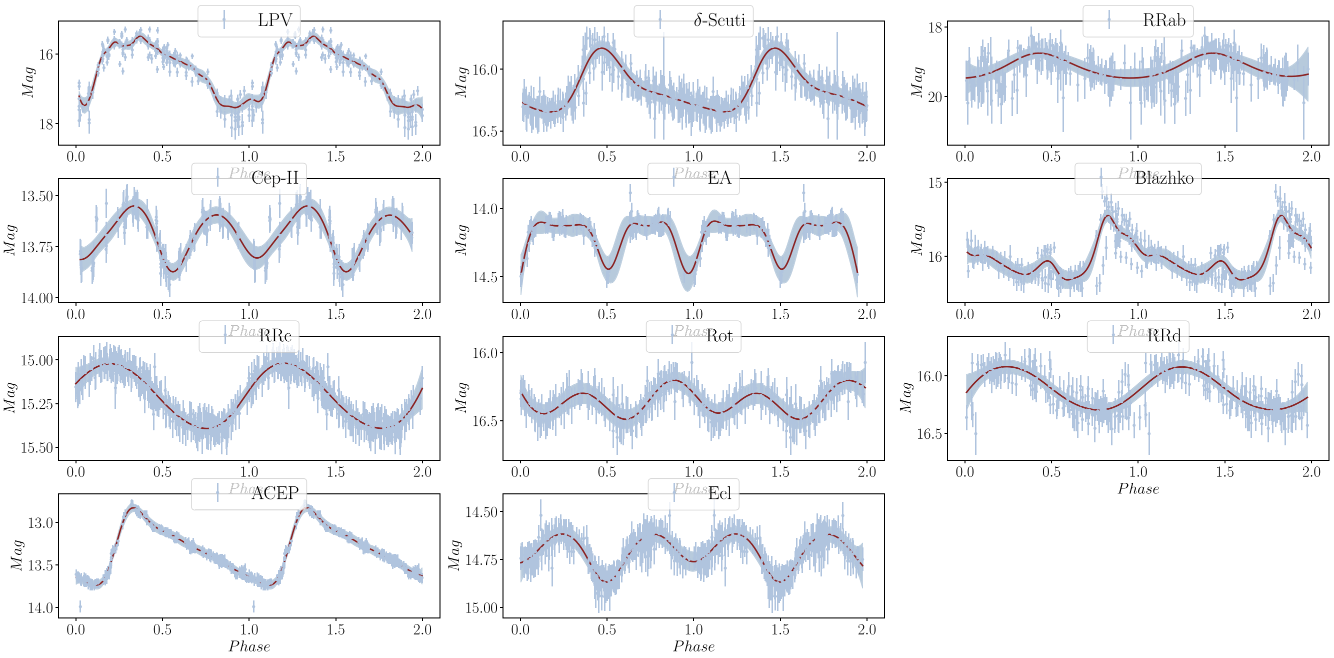

Figure 1. Folded light curves of different kinds of variable stars reported in magnitudes as a function of phase. The data points are represented in light blue dots along with the error bars, and the fitted light curves are illustrated in purple lines.

1. It can get better classification performance.
2. It uses less training data.
3. For sampling, it requires less computation time.
4. It is robust for data with noise or missing values.
5. Compared to traditional imbalanced learning, SPE is less influenced by hyperparameters since it belongs to ensemble learning methods.
6. It provides various base classifiers for choice.
7. It does not depend on the distance metric and can also be applied for discrete data without modification.

The SPE algorithm introduces the concept of 'classification hardness distribution', which reflects the task difficulty related to factors such as noise, model capacity, and class imbalance. Instinctively, hardness means the difficulty of accurately classifying a sample with a classifier. So hardness distribution is helpful to guide the re-sampling strategy to obtain better performance. Rather than simply balancing dataset or directly assigning class

https://doi.org/10.1017/pasa.2023.35 Published online by Cambridge University Press

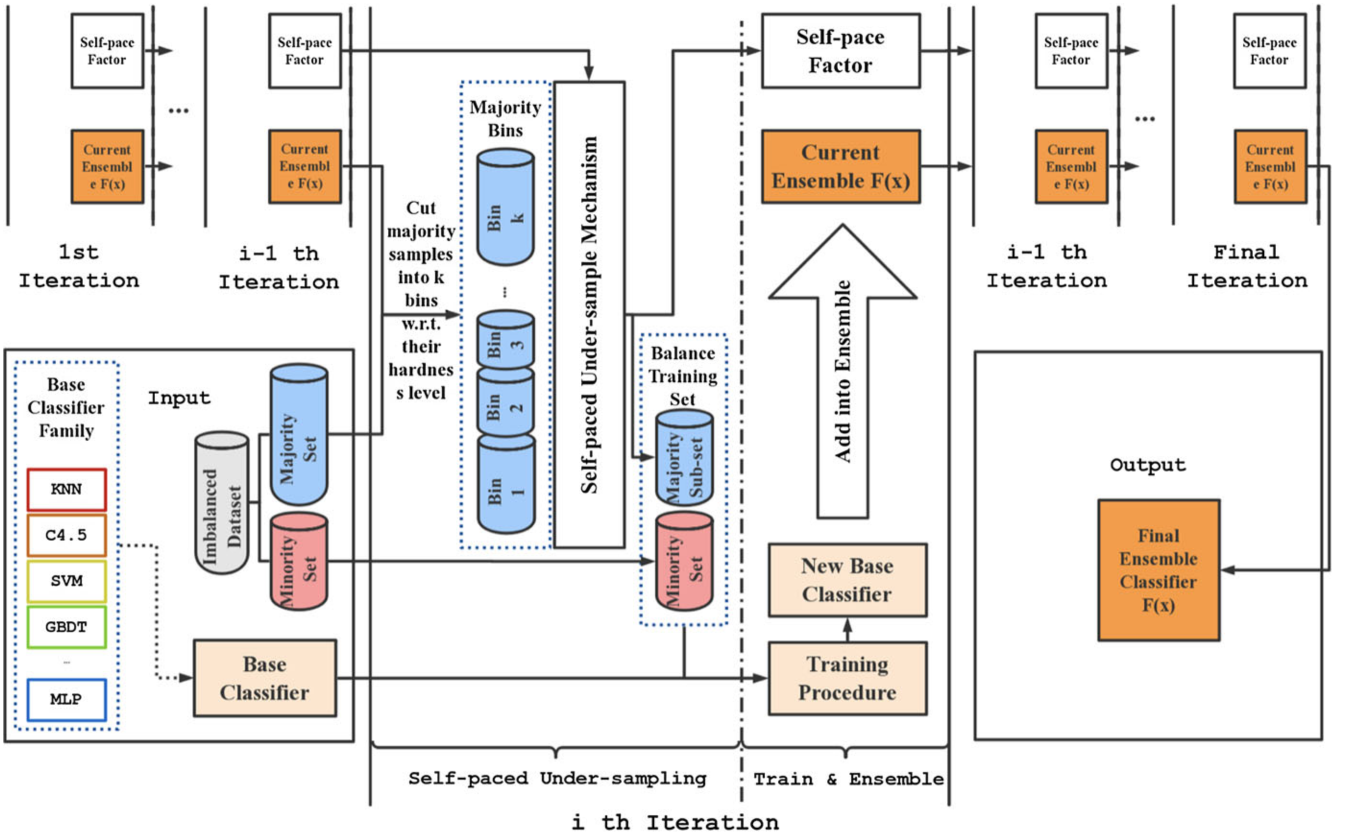

Figure 2. The pipeline of Self-paced Ensemble from Liu et al. (2020). Instead of simply balancing the data or directly adjusting class weights, classification hardness is taken into account over the dataset, and the most informative majority data samples are iteratively selected according to the hardness distribution. The under-sampling strategy is controlled by a self-paced procedure, which enables SPE to gradually focus on the harder data samples but still retains the information of the majority sample to prevent over-fitting.

weights, the distribution of classification hardness is taken into account, and SPE iteratively selects the most informative majority datasets based on the hardness distribution. Boosting-like serial training is performed using under-sampling and ensemble strategy, and finally an additive model is obtained. Specifically, the under-sampling is controlled by a self-paced procedure, which makes the structure gradually concentrate on the harder samples (i.e., minority sample). All the majority samples are split into k bins based on their hardness values. To harmonise the hardness contribution of each bin, the sample probability of those bins gradually decreases for the majority samples and the declining level is controlled by a self-paced factor. In the first few iterations, the framework primarily focuses on informative samples. In the later iterations when a self-paced factor becomes very large, the information of the majority sample is still retained to prevent over-fitting. So SPE is an efficient, general-purpose, and robust ensemble imbalanced learning algorithm. Now SPE is a part of imbalanced-ensemble toolbox, built on the basis of both scikit-learn (Pedregosa et al. 2011) and imbalanced-learn.[b] So SPE is directly utilised from the imbalanced-ensemble package in our work.

3.2. Feature extraction

In time-domain astronomy, the data collected from telescopes are usually expressed in the form of light curves. These light curves show the brightness changes of stars over a period of time. The extraction of light curve features is a part of this work, which can be used to characterise and distinguish different variables. Features can range from basic statistical attributes such as mean and standard deviation, to more complex time series features such as autocorrelation functions. Ideally, these features should be informative and discriminative, enabling machine learning algorithms to distinguish the categories of light curves. FATS (Nun et al. 2015) is used for feature extraction. We select the features that can best capture the properties of the light curves for imbalanced classification of periodic variable stars.

We choose seven features for classification. They are mean magnitude, standard deviation, mean variance, skew, kurtosis, amplitude, and period. Six of these features are computed by FATS, and the period is from the downloaded catalogue. These features reflect the location, scale, variability, morphology, and observation time of the light curves. They are easy to interpret and robust against bias.

3.3. Hyperparameter optimisation

In previous works, researchers usually used the greedy grid search for hyperparameter optimisation. Hutter, Hoos, & Leyton-Brown (2011) found that Bayesian optimisation, also called sequential model-based optimisation (SMBO), outperformed grid search for large parameter spaces. Compared to SMBO, grid search has slower speed and more computation cost. Moreover, it is affected by setting range of each hyperparameter. In reality, it is impossible

[b] https://imbalanced-learn.readthedocs.io/en/stable/index.html.

https://doi.org/10.1017/pasa.2023.35 Published online by Cambridge University Press

Table 3. Confusion matrix of binary classification.

Label Predict	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

to try all possible values of hyperparameters. Considering these factors, we adopt SMBO for selecting optimal hyperparameters. In this work, SMBO is combined with the SPE algorithm to establish a better classifier at high speed.

3.4. Evaluation metric

For imbalanced learning, the accuracy is not a good measure of the performance of a classifier. So, we usually use other evaluation metrics based on true positive (TP), false negative (FN), false positive (FP), and true negative (TN). For the binary classification, they can be recorded in a confusion matrix, as shown in Table 3. For evaluating the performance of algorithms, *Recall* and *Precision* are commonly used. For imbalanced datasets, we also consider *Balanced Accuracy*, $G-Mean$ (harmonic or geometric mean of *Precision* and *Recall*; García, Sánchez, & Mollineda 2007), and AUCROC (the area under receiver operator characteristic curve; Sahiner et al. 2017).

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Specificity = \frac{TN}{FP + TN} \tag{3}$$

$$Balanced\ Accuracy = \frac{Recall + Specificity}{2} \tag{4}$$

$$G\ Mean = \sqrt{Recall \times Precision} \tag{5}$$

4. Results and discussion

Our aim is to classify periodic variables based on the CRTS database. Sample imbalance and the performance of a classifier may influence classification results. Our work involves the imbalance problem. Here, we plan to apply the SPE algorithm and Voting Classifier to solve it.

The samples are randomly split into training and test sets for 10 times, with a ratio of 7:3. For the SPE algorithm, it can be used to boost any canonical classifier's performance (e.g., SVM, C4.5 Quinlan 1986, Random Forest, or Neural Networks Haykin 2009). Random Forest is selected as the base classifier in our work. Because the number of base classifiers (*n_estimators*) significantly influences the performance of ensemble methods, *n_estimators* should be tuned. We use SMBO to select the optimal number of base classifiers.

4.1. Comparing SPE with hierarchical tree classifier

Ho20 proposed a hierarchical tree classifier (HTC) with three layers for periodic variables using the CRTS catalog. They argued that simple under-sampling methods were not enough to deal with the class imbalance problem. Instead, they applied GpFit (Williams & Rasmussen 2006) to generate artificial data and increase the number of training samples. They used GpFit for data augmentation and Random Forest as the base algorithm for each layer of the HTC. This approach combined data-level and algorithm-level solutions for imbalanced learning. In contrast, our work uses a self-paced under-sampling procedure that gradually focuses on the harder samples and prevents over-fitting. We train and test the SPE algorithm directly on the original data without building a HTC or augmenting the data. To compare the performance of Ho20's method and our SPE method, we use AUCROC as a comprehensive metric that reflects the trade-off between *sensitivity* and *specificity*. We find that Ho20's method performs poorly for classes with small sample sizes, such as Rotational, RRd, Type-II Cepheid, and Blazhko RR Lyrae variables. Fig. 3 shows that our SPE method significantly improves the performance for Blazhko RR Lyrae variables but slightly worse for RRc variables. We also use confusion matrices to compare the methods, as shown in Figs. 4–5. For Ho20's method, the HTC divides all variables into Eclipsing, Rotational, and Pulsating variables in the first layer, with *Recall* values of 72%, 67%, and 88%, respectively. These values are not very high, and any misclassification in this layer will propagate to the next layers. For the second layer, the classification of Eclipsing and Pulsating variables seems good, but it may be influenced by the previous layer. For the third layer, Fig. 5 shows poor performance for separating RRd and Blazhko variables. In fact, for Ho20's method, misclassification accumulates from the top layer to the bottom layer. Therefore, the final classification metrics should be computed by multiplying the metrics obtained in three layers. Compared to Ho20's method, our SPE method improves the classification *Recall* of Blazhko RR Lyrae stars from 12% to 85% and of mixed-mode RR Lyrae variables from 29% to 64%; for detached binaries, the classification *Recall* increases from 68% to 97%; for LPV, the classification *Recall* rises from 87% to 99%. The only exceptions are RRab, RRc, and contact and semi-detached binary classes.

There are some evaluation metrics for imbalanced problems. Here, we consider *Balanced Accuracy* and *G Mean*. Tables 4–5 represent *Balanced Accuracy* and *G Mean* for SPE and the method of Ho20. We observe that our SPE method increases *Balanced Accuracy* of Blazhko stars from 11% to 60% and of RRd stars from 24% to 60%. However, this improvement comes with a slight decrease in the classification of RRab and RRc subtypes. For detached binaries, our SPE method improves *Balanced Accuracy* from 52% to 87% and *G Mean* from 75% to 93%. For rotational variables, our SPE method performs worse than Ho20's method, because we separate rotational variables from all the subtypes in one step, while Ho20's method uses a three-layer HTC. If we use a similar HTC as Ho20, *Recall* of Eclipsing, Rotational, and Pulsating variables by our SPE method is 69%, 73%, and 89%, respectively. *Recall* (73%) of Rotational variables by our SPE method is higher than that (67%) of Ho20's method. Therefore, our approach significantly improves the classification performance compared to Ho20's work and achieves good results for most classes of variables except RRab, RRc, and contact and semi-detached binary.

To summarise, the SPE algorithm can classify all the variables in one step without using HTC, and its ability to increase *Recall* makes it desirable. The results show that the SPE algorithm favours the small sample classes, such as RRd, Blazhko, Delta-scuti, and

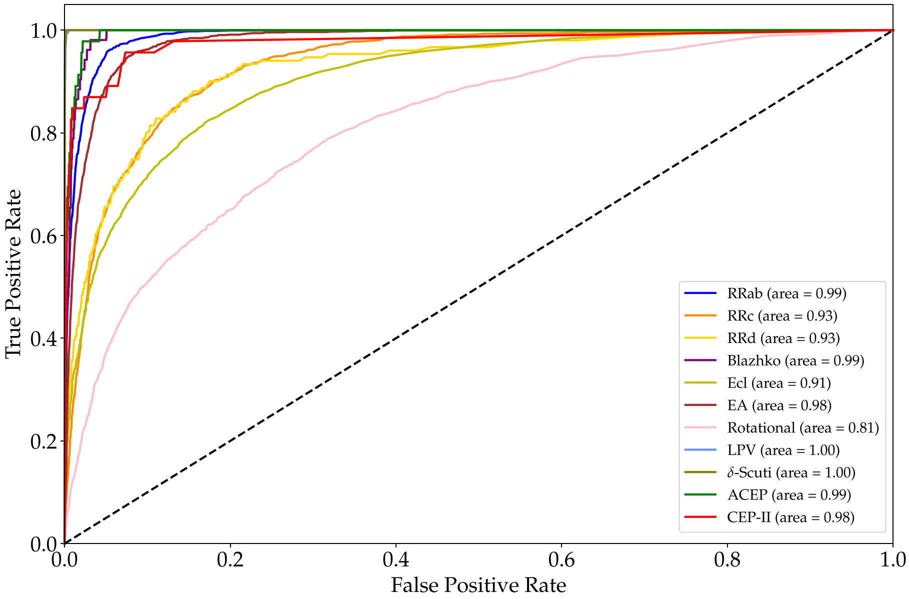

Figure 3. The AUCROC of SPE for different classes.

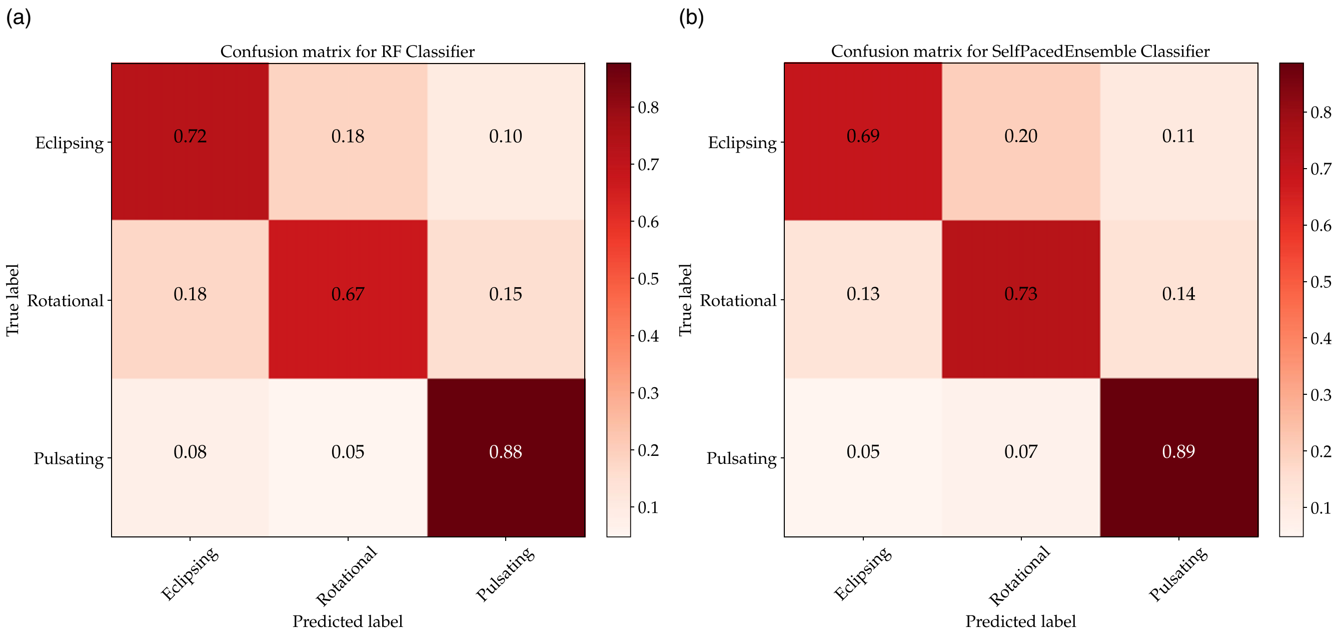

Figure 4. Comparison of confusion matrix of SPE with the first-layer classifier of Ho20.

Cepheid stars, and achieves high classification performance for them. However, the base algorithm of HTC is Random Forest, which is similar to other traditional classifiers and tends to select the majority classes, such as RRab and RRc stars. Therefore, there is a trade-off between SPE and Random Forest. To balance their performance, we will use a Voting Classifier that combines SPE and Random Forest in Section 4.2.

4.2. *Voting Classifier*

The Voting Classifier is a method that combines different machine learning classifiers and uses a majority vote or the average predicted probabilities (soft vote) to predict the class labels. We use soft voting and return the class label as argmax of the average of predicted probabilities, as explained in the website[c] and Fig. 6. The voting mechanism may help us to balance the performance between SPE and Random Forest and thus improve the classification accuracy. We build Random Forest and SPE models separately and then use voting to return the class label based on the average predicted probabilities of both models. Figs. 7 and 8 show that Voting Classifier can correct errors made by any individual classifier, leading to better performance. Specifically, Voting Classifier finds a balance between Random Forest and SPE.

[c] https://scikit-learn.org/stable/modules/ensemble.html#voting-classifier.

https://doi.org/10.1017/pasa.2023.35 Published online by Cambridge University Press

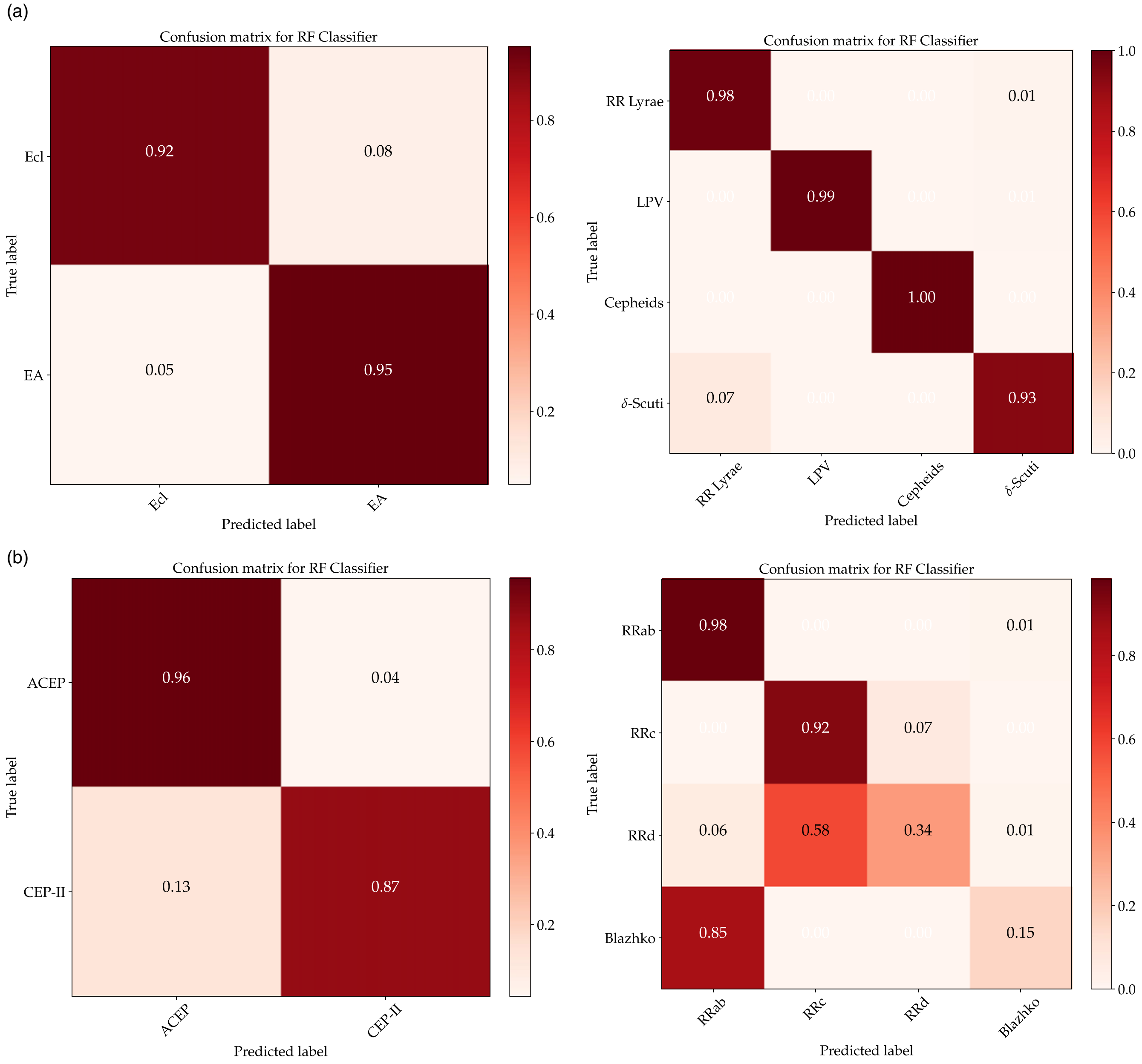

Figure 5. Confusion matrix of the second-layer and third-layer classifier in Ho20.

Compared to the SPE classifier, Voting Classifier increases the classification *Recall* of the RRab and RRc stars but decreases it for the RRd and Blazhko stars. Compared to Random Forest, Voting Classifier retains some advantages of the SPE classifier. So it can also be sensitive to small sample classes, such as Delta-scuti and Cepheid stars. We also compute *Balanced Accuracy* for Random Forest, SPE, and Voting Classifier, which are 71%, 78%, and 77%, respectively. Table 6 shows that Voting Classifier can balance the performance of classifiers. Therefore, for different research goals, we may adopt different strategies. For large surveys, a comprehensive classifier is necessary, and Voting Classifier is useful to obtain a balanced classification. If we are interested in the small classes, such as Blazhko and Delta-Scuti stars, we may use the SPE classifier. Similarly, if we tend to select the majority classes, we may use the Random Forest algorithm.

4.3. Some factors affecting the performance of a classifier

Since class imbalance is not the fundamental cause of classification difficulty (Liu et al. 2020), there are other factors influencing the performance of a classifier, as follows:

1. Some minority class samples appear in the distribution of dense majority class samples.
2. Overlapping between classes (García et al. 2007).

https://doi.org/10.1017/pasa.2023.35 Published online by Cambridge University Press

Table 4. Mean *Balanced Accuracy* and *G Mean* for SPE.

Classes of variable stars	*Balanced Accuracy*	*G Mean*
RRab	0.69	0.84
RRc	0.58	0.77
RRd	0.60	0.78
Blazhko	0.84	0.92
Contact & Semi-Detached Binary	0.49	0.71
Detached Binary	0.87	0.93
Rotational Variable	0.48	0.70
Long Period Variable	0.99	1.00
Delta-Scuti	1.00	1.00
Anomalous Cepheid	0.85	0.93
Type-II Cepheid	0.78	0.89

Table 5. Mean *Balanced Accuracy* and *G Mean* for the work of Ho20.

Classes of variable stars	*Balanced Accuracy*	*G Mean*
RRab	0.72	0.85
RRc	0.66	0.81
RRd	0.24	0.51
Blazhko	0.11	0.34
Contact & Semi-Detached Binary	0.55	0.75
Detached Binary	0.52	0.75
Rotational Variable	0.55	0.75
Long Period Variable	0.78	0.89
Delta-Scuti	0.72	0.85
Anomalous Cepheid	0.66	0.81
Type-II Cepheid	0.64	0.81

3. Minority class is split into small disjuncts due to sparsity, which is abbreviated as small disjuncts (Prati, Batista, & Monard 2004).

In our work, overlapping between classes makes it hard for a classifier to separate the minority from the majority. For example, subtypes of RR Lyrae stars overlap because they have no clear physical differences. RRab stars pulsate in fundamental mode while RRc stars pulsate in the first overtone, so they can be separated well. However, RRd stars pulsate in both modes, so they are tricky to distinguish from RRab and RRc stars. Furthermore, Blazhko effect occurs among RRab, RRc, and RRd stars. Fig. 9 shows that it is not easy to separate RRc stars from RRd stars, and RRab from Blazhko stars. Only for show, the scatter plot of Period versus Smallkurtosis is given here, actually it is difficult to distinguish different RR Lyrae stars by any pairwise combination of features due to their overlapping. Rotating variables show small luminosity changes from patches of light spots on their surfaces, and they may have bright spots at the magnetic poles. Moreover, they often belong to binary systems. All these facts lead to the confusion of rotating stars with other classes.

Another reason is that the labels of periodic variables depend on the classification of existing sky surveys or different experts. The classifier's performance will suffer from the wrong labels. If the standard variable star sets or accurate labels are provided, the classifier will obtain a more reliable performance.

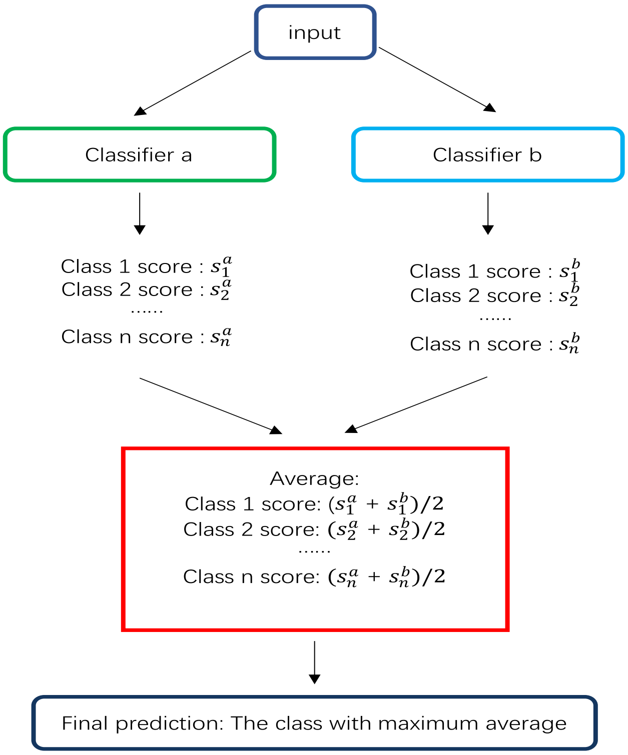

Figure 6. The flowchart of the Voting Classifier. In Soft voting, classifiers or base models are fed with training data to predict the class output of *n* possibilities. Each classifier independently assigns the occurrence probability of each class. In the end, the average of the possibilities of each class is calculated, and the final output is labelled as the class with maximum average.

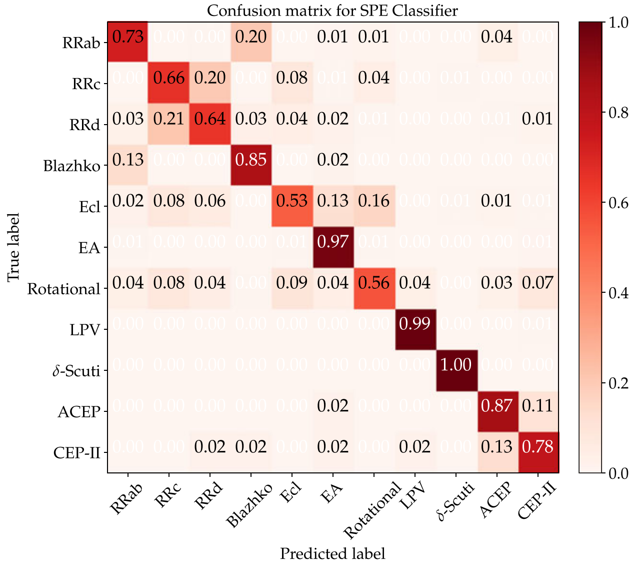

Figure 7. Confusion matrix of SPE.

As a result, the SPE algorithm is suitable for the classification task of targeting minority samples, compared to traditional imbalanced learning. By combining SPE with Random Forest in a Voting Classifier, we can achieve better balanced classification performance.

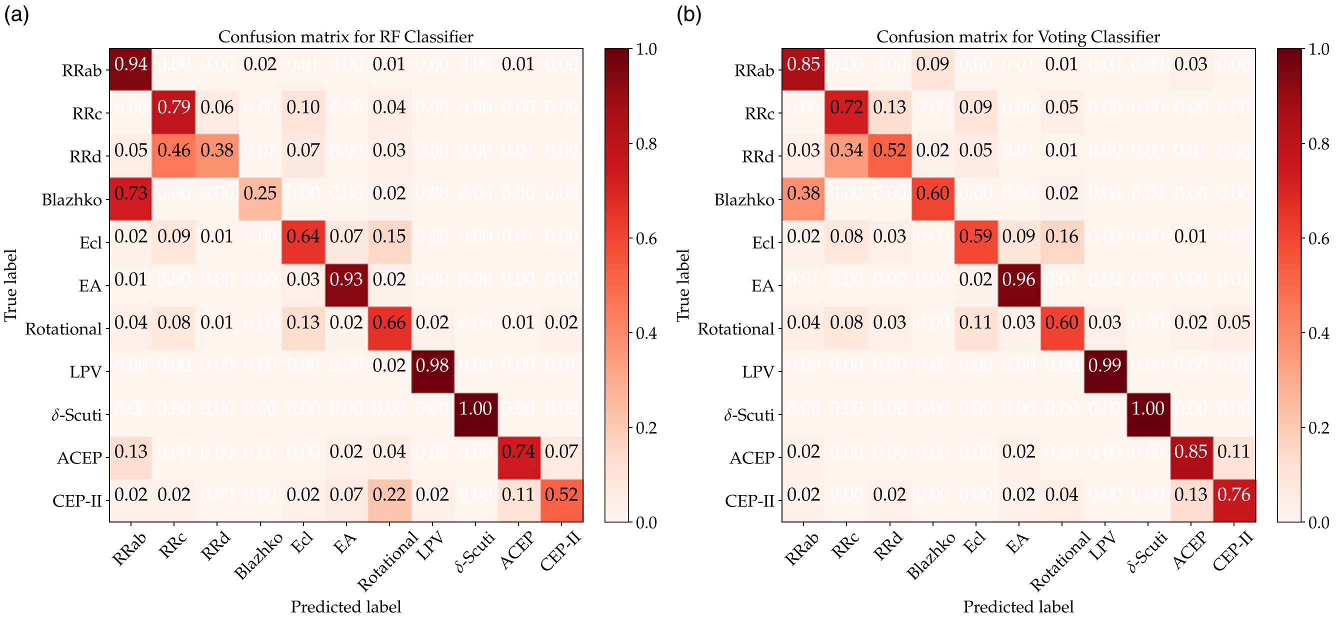

Figure 8. A Voting Classifier is built by combining Random Forest and SPE. As shown in Fig. 7, the left and right panels depict the confusion matrices of Random Forest and Voting Classifier, respectively.

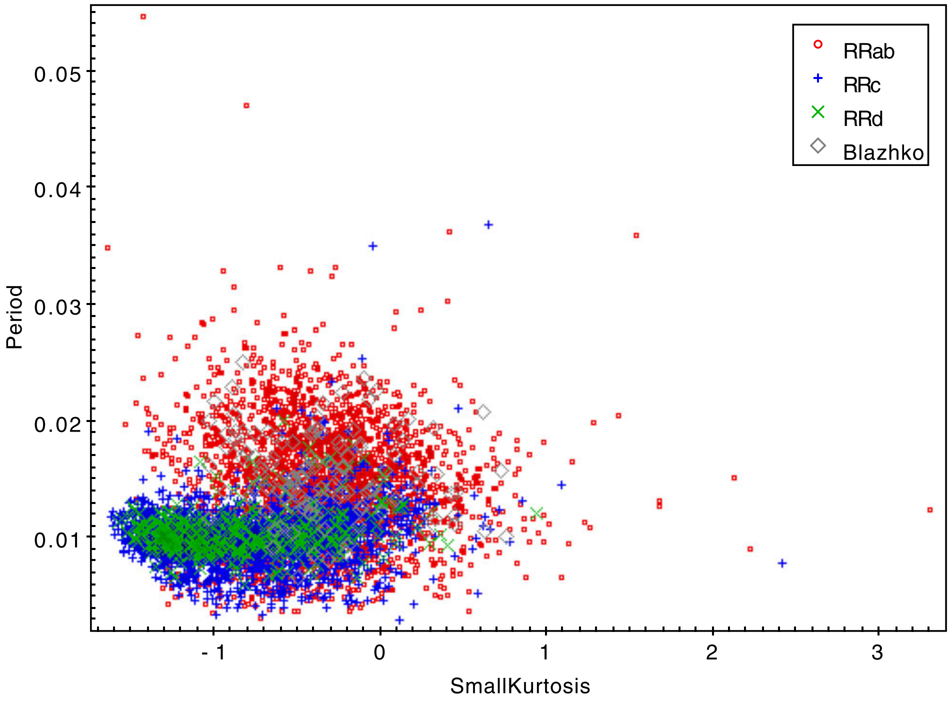

Figure 9. The Period versus Smallkurtosis distribution of RRab, RRc, RRd, and Blazhko stars.

5. Conclusion

In this work, we use the SPE algorithm and Voting Classifier to classify different variables and address the imbalanced classification problem. Compared to Ho20's work, which uses GpFit for data augmentation and Random Forest for hierarchical classification, our SPE method uses the original dataset without data processing and classifies all variables in one step. Moreover, our SPE method avoids the propagation of misclassification from one layer to another in the hierarchical trees. Therefore, our SPE method is better when these factors are considered. The results also show that the SPE algorithm improves the performance for minority classes at the expense of some majority classes without

https://doi.org/10.1017/pasa.2023.35 Published online by Cambridge University Press

Table 6. Mean *Balanced Accuracy* and *G Mean* for Voting Classifier.

Classes of variable stars	*Balanced Accuracy*	*G Mean*
RRab	0.82	0.91
RRc	0.66	0.82
RRd	0.48	0.71
Blazhko	0.57	0.77
Contact & Semi-Detached Binary	0.54	0.75
Detached Binary	0.89	0.94
Rotational Variable	0.51	0.73
Long Period Variable	0.99	1.00
Delta-Scuti	1.00	1.00
Anomalous Cepheid	0.83	0.92
Type-II Cepheid	0.74	0.87

losing the overall accuracy. Considering the trade-off between SPE and Random Forest, we use the Voting Classifier to balance the overall classification performance. In practice, the overlapping of classes and inconsistent labels may lead to misclassification and affect the performance of classifiers. Therefore, a complete and representative known sample is essential for building an excellent classifier. However, such samples are scarce, especially for rare objects. In this case, the SPE algorithm shows its superiority. For the identification of minority classes, the SPE algorithm and Voting Classifier are efficient and reliable, and they can be applied to the time-domain data of other larger sky survey projects (LSST, etc.).

Acknowledgements. We are grateful to the anonymous referee for the constructive comments, which significantly helps improve our paper. This paper is funded by the National Natural Science Foundation of China under grants Nos.12203077, No.12273076, No.12133001, and No.11873066, the science research grants from the China Manned Space Project with Nos. CMS-CSST-2021-A04 and CMS-CSST-2021-A06, and Natural Science Foundation of Hebei Province No.A2018106014. We acknowledgement the CRTS databases.

Data availability. Variables of the Catalina Surveys DR2 are available from http://nesssi.cacr.caltech.edu/DataRelease/VarcatS.html.

References

Alloin, D., & Gieren, W. 2003, Stellar Candles for the Extragalactic Distance Scale, Vol. 635, doi: 10.1007/b13985
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, JAIR, 16, 321
Chen, X., Wang, S., Deng, L., de Grijs, R., & Yang, M. 2018, ApJS, 237, 28
Drake, A. J., et al. 2017, MNRAS, 469, 3688
Eyer, L., & Mowlavi, N. 2008, in Journal of Physics Conference Series, Vol. 118, Journal of Physics Conference Series, 012010
Gao, D., Zhang, Y.-X., & Zhao, Y.-H. 2009, RAA, 9, 220
García, V., Sánchez, J. S., & Mollineda, R. A. 2007, in Congress on Progress in Pattern Recognition, Image Analysis Applications
Graham, M., Drake, A., Djorgovski, S. G., Mahabal, A., & Donalek, C. 2017, in European Physical Journal Web of Conferences, Vol. 152, European Physical Journal Web of Conferences, 03001
Haykin, S. 2009, Neural Networks and Learning Machines, 3/E (Pearson Education India)
Hosenie, Z., Lyon, R., Stappers, B., Mootoovaloo, A., & McBride, V. 2020, MNRAS, 493, 6050
Hoyle, B., et al. 2015, MNRAS, 452, 4183
Hutter, F., Hoos, H. H., & Leyton-Brown, K. 2011, in International Conference on Learning and Intelligent Optimization (Springer), 507
Jin, X., et al. 2019, MNRAS, 485, 4539
Koch, D. G., et al. 2010, ApJ, 713, L79
Koposov, S. E., et al. 2019, MNRAS, 485, 4726
Liu, Z., et al. 2020, in 2020 IEEE 36th International Conference on Data Engineering (ICDE) (IEEE), 841
Nun, I., et al. 2015, arXiv e-prints, arXiv:1506.00010
Pedregosa, F., et al. 2011, JMLR, 12, 2825
Peng, N., Zhang, Y., & Zhao, Y. 2013, SCPMA, 56, 1227
Petrosky, E., Hwang, H.-C., Zakamska, N. L., Chandra, V., & Hill, M. J. 2021, MNRAS, 503, 3975
Prati, R. C., Batista, G. E., & Monard, M. C. 2004, in Brazilian Symposium on Artificial Intelligence (Springer), 296
Price-Whelan, A. M., et al. 2019, AJ, 158, 223
Prudil, Z., Dékány, I., Grebel, E. K., & Kunder, A. 2020, MNRAS, 492, 3408
Quinlan, J. R. 1986, ML, 1, 81
Sahiner, B., Chen, W., Pezeshk, A., & Petrick, N. 2017, in Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment, Vol. 10136 (SPIE), 100
Williams, C. K., & Rasmussen, C. E. 2006 (MA: MIT Press Cambridge)
Wright, E. L., et al. 2010, AJ, 140, 1868
Zhang, J., Zhang, Y., & Zhao, Y. 2018, AJ, 155, 108
Zhang, J., Zhang, Y., & Zhao, Y. 2020, ApJS, 246, 8
Zhang, Y., Zhao, Y., & Wu, X.-B. 2021, MNRAS, 503, 5263

https://doi.org/10.1017/pasa.2023.35 Published online by Cambridge University Press