RAA 2021 Vol. **21** No. **4**, 99(11pp) doi: 10.1088/1674-4527/21/4/99

http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

The feasibility and flexibility of selecting quasars by variability using ensemble machine learning algorithms

Da-Ming Yang (羊达明), Zhang-Liang Xie (谢彰亮) and Jun-Xian Wang (王俊贤)

CAS Key Laboratory for Researches in Galaxies and Cosmology, University of Science and Technology of China, Chinese Academy of Sciences, Hefei 230026, China; *ydm2016@mail.ustc.edu.cn*
School of Astronomy and Space Science, University of Science and Technology of China, Hefei 230026, China

Received 2020 August 27; accepted 2020 November 4

Abstract In this work, we train three decision-tree based ensemble machine learning algorithms (Random Forest Classifier, Adaptive Boosting and Gradient Boosting Decision Tree respectively) to study quasar selection in the variable source catalog in SDSS Stripe 82. We build training and test samples (both containing 1:1 of quasars and stars) using the spectroscopic confirmed sources in SDSS DR14 (including 8330 quasars and 3966 stars). We find that when trained with variation parameters alone, all three models can select quasars with similarly and remarkably high precision and completeness ($\sim$ 98.5% and 97.5%), even better than trained with SDSS colors alone ($\sim$ 97.2% and 96.5%), consistent with previous studies. By applying the trained models on the variable sources without spectroscopic identifications, we estimate the spectroscopically confirmed quasar sample in Stripe 82 variable source catalog is $\sim$ 93% complete (95% for $m_i < 19.0$). Using the Random Forest Classifier we derive the relative importance of the observational features utilized for classifications. We further show that even using one- or two-year time domain observations, variability-based quasar selection could still be highly efficient.

Key words: quasars: general — catalogs — methods: data analysis

1 INTRODUCTION

Quasars, as one of the most luminous celestial objects in the universe, are powered by accreting supermassive black holes (SMBHs) in galactic nuclei. Building quasar catalogs is of great significance to research on SMBH accretion, galaxy evolution and cosmic structure. Since spectroscopic observations are always expensive in observing time, pre-selecting quasar candidates from large area photometric observations is an essential topic in this area. To date, the largest samples of quasars have been selected based on their optical/UV colors which are often different from those of stars (e.g. Richards et al. 2002; Pâris et al. 2018; Yao et al. 2019). Infrared colors are also useful to select active galaxies and quasars (e.g. Stern et al. 2005; Lacy et al. 2007; Donley et al. 2012). Meanwhile, quasar candidates can be pre-selected based on multi-band detections, e.g., in X-ray, or in radio band.

In the era of time domain astronomy, selecting quasar candidates based on variation properties emerges as a new frontier. Flux variation in multi-band is one of the most prominent characteristics of quasars and active galaxies. In optical, quasars are often more variable than stars (Sesar et al. 2007), and studies have shown that quasars can be efficiently selected based on their optical variation properties (e.g. MacLeod et al. 2011; Choi et al. 2014).

Variation-based quasar selection from large area time domain surveys could be uniquely helpful since it is free from the known biases suffered by the traditional color selection approach (e.g. Butler & Bloom 2011; Sánchez-Sáez et al. 2019). For example, the traditional optical/UV color selection is insensitive to quasars at the redshift range of 2.2 to 3 in which their colors are similar to that of normal quasars (e.g. Schneider et al. 2010).

While quasars can be selected using empirical cuts in the space of variation parameters (e.g. MacLeod et al. 2011; Schmidt et al. 2010; Butler & Bloom 2011), machine learning algorithms have been adopted by several studies to improve the precision and completeness of the selection. In these studies, various algorithms have been trained using various datasets. Graham et al. (2014) used Slepian wavelet variance (SWV), Damped Random Walk (DRW) and Structure Function (SF) at the same time to extract variability features from Catalina Real-time Transient

Survey (CRTS) quasar samples, along with color features to train different ensemble learning classifiers, including Random Forest, Extremely randomized trees, AdaBoost and Gradient tree boosting. They also used SDSS Stripe 82 data to test their sample selection criteria. Hernitschek et al. (2016) utilized color and variability features and trained a Random Forest Classifier (RFC) to identify quasars on Pan-STARR1 3π survey, in which each quasar was observed typically seven times in each of its five bands within 3.5 years. Similarly, Sánchez-Sáez et al. (2019) used RFC to classify AGNs in QUEST–La Silla AGN variability survey. Takata et al. (2018) employed Support Vector Machine (SVM) to classify variable quasars in Sloan Digital Sky Survey Stripe 82, and Kim et al. (2011) also adopted SVM to identify variable quasars in MAssive Compact Halo Object (MACHO) database.

While these works have shown that the machine learning selection of quasars by variability is a promising approach, it would be useful to investigate how the performance of the machine learning approach depends on the availability of bands and the length of the light curves from which the variability parameters are measured. In this work, we present an extensive study of machine learning selection of quasars by variability. We use the SDSS Stripe 82 variable source catalog to train three decision-tree based machine learning algorithms, including RFC, Adaptive Boosting (AdaBoost, Freund & Schapire 1995) and Gradient Boosting Decision Tree (GBDT, Friedman 2002; Mason et al. 1999). The ground-truth labels of objects we use in this work come from SDSS Data Release 14 (DR14) spectroscopic database. In Section 2, we introduce the dataset we used. In Section 3, we describe the variability features and the training procedure, including the optimal hyper-parameters of the ensemble machine learning methods. In Section 4.1, we confirm that the three machine learning classifiers could yield remarkably high precision and completeness in classifying quasars. In Section 4.2 we apply the trained algorithms to classify variable sources without spectroscopic identifications, and estimate the completeness of the quasar sample in SDSS Stripe 82. We present and discuss the relative importance of the observational features (variability, color, from various bands) we utilize in Section 4.3. The effects of imbalanced samples (the ratios of quasars and stars used in training and test samples are not equal to 1) are discussed in Section 4.4. We explore the dependence of quasar classification on the length of light curves in Section 4.5. Finally, our brief conclusions are given in Section 5.

2 SDSS STRIPE 82 VARIABLE SOURCES

SDSS Stripe 82 is a 290 deg^2 equatorial field, which has been repeatedly scanned $\sim$ 60 times in $ugriz$ within $\sim$ 10 years by the Sloan Digital Sky Survey (Ivezić et al. 2007). A catalog of 67 507 variable sources in SDSS Stripe 82 was built by Ivezić et al. (2007). Thanks to the long duration and the large number of visits for each source, variation parameters could be measured with considerably high accuracy. The catalog, together with the SDSS spectroscopical identifications, is uniquely useful to promote systematical study and understanding of variation-based quasar classification. This study could be essential to quasar selection from upcoming large area time domain surveys (e.g. LSST).

This catalog was built following the criteria listed below:

- unresolved in imaging data, with photometric error below 0.05 mag in at least one band;
- processing flags BRIGHT, SATUR, BLENDED, or EDGE are not set in any band;
- at least 10 photometric observations in the g and r bands;
- the median g-band magnitude brighter than 20.5;
- root-mean-square scatter $>$ 0.05 magnitude and χ^2 per degree of freedom larger than three in both g and r bands, which mean the variation is statistically significant (see the discussions in Sesar et al. 2007 for the details).

We match the catalog with a matching radius of $2''$ with SDSS Data Release 14 (DR14, Abolfathi et al. 2018) for spectroscopical identifications. Among the variable sources, 8330 are identified as quasars and 3966 as stars. The rest 48 716 sources are left unlabeled. The spectroscopically identified sources are utilized to train and test our machine learning classifiers, which are also utilized to classify those unlabeled sources.

3 OBSERVATIONAL FEATURES AND MACHINE LEARNING MODELS

3.1 Observational Features

We use a DRW (also named as Ornstein-Uhlenbeck process) process to fit the light curves. The DRW process is a stochastic process defined by an exponential covariance matrix between t_i and t_j:

$$S_{\rm DRW}(\Delta_t) = \sigma^2 \exp(-|\frac{\Delta_t}{\tau}|) \quad (1)$$

where σ is the long-term deviation of variability, and τ the characteristic time scale of DRW. It is essentially a random walk with a self-correcting term that pushes any deviations back toward the mean value of the random walk itself. Various studies have shown that the DRW process could well describe the observed UV/optical variations

of active galaxies and quasars (e.g. Kelly et al. 2009; Kozłowski et al. 2010; MacLeod et al. 2010) [1]. Compared with many other deterministic and stochastic models, the DRW process is the best model for SDSS Stripe 82 quasar light curves (Andrae et al. 2013). Meanwhile, the DRW parameters had been found useful to distinguish quasars from variable stars. For instance, MacLeod et al. (2011) found that the distribution of τ (in the observed frame) peaks around 500 days for quasars in Stripe 82, but $\sim$ 1 day for other objects, showing selecting quasars out of variable sources is highly promising even without color information.

In this work, we use the DRW parameters as the input observational features to train and test our machine learning classifiers. We use the software JAVELIN (Zu et al. 2011, 2013) to fit each SDSS Stripe 82 light curve to measure τ and σ. A total of 10 DRW parameters (for five SDSS bands) were obtained for each variable source. The DRW fitting failed for two stars and 110 unlabelled sources, which are excluded from further analyses. We also note some stars with "peculiar" fitted DRW parameters, which however do not affect the analyses in this work. An example light curve of such sources is presented in Appendix A, which results in an unreasonably large σ. We presume that the DRW fitting may fail or yield unreasonable results owing to the fact that the variability of such sources is far from a DRW process.

For light curves with a small number of epochs, the DRW fitting might yield parameters with huge uncertainties. Therefore, we also measure the maximum variation amplitudes of each source (in gri bands, which have considerably better SDSS photometry comparing with u and z), and include them as input to the classifiers. The relative importance of these input features are presented and discussed in Section 4.3.

It would be interesting to examine, for the variation selected sources in this study, whether the variation features alone can better classify quasars comparing with colors, and whether combining variation and color features can further improve the classification. The color features of the variable sources ($u-g$, $g-r$, $r-i$ and $i-z$) are thus also considered, which were calculated using the median photometry in each band.

[1] Though deviations from DRW at extremely short and extremely long timescales have been reported (e.g. Mushotzky et al. (2011), Zu et al. (2013), Guo et al. (2017)). More sophisticated models include mixture of Ornstein-Uhlenbech (OU) processes (Kelly et al. 2011), continuous auto regression and moving average (CARMA) model (Kelly et al. 2014; Simm et al. 2016; Kasliwal et al. 2017), and a broken powerlaw shaped power density spectrum (e.g. Zhu & Xue 2016) have been introduced.

3.2 Machine Learning Models

We use RFC as our major classifier in this study. RFC is one of the most popular and powerful supervised machine learning methods. It belongs to a subclass called ensemble learning method, which includes many weak classifiers, and all of the weak classifiers collaboratively make a final decision. In the case of random forest, it includes hundreds (or even thousands) of decision trees, each of them is trained by the training set. An example of decision tree trained in this study is presented in Figure 1.

Each internal node stands for a selection rule on a specific feature, which splits the node into two branches. Each leaf node (nodes at the bottom of a tree) stands for a class, in our case, a quasar or a star. In real training, a branch of a decision tree will stop growing deeper once the purity of the newest node reaches the preset value (not necessary to be 100%) , and the node will become a leaf node.

We also adopt two other decision-tree-based models, namely: Adaptive Boosting (AdaBoost, Freund & Schapire 1995) and Gradient Boosting Decision Tree (GBDT, Friedman 2002; Mason et al. 1999). Unlike random forest, these models use boosting but not bagging strategy (random forest is essentially a bagging ensemble learning model). AdaBoost is adaptive because the misclassified samples will be used specifically in the next iteration, and the outputs of decision trees ("weak learners") in each epoch are combined into a weighted sum as a final output. For GBDT, in each iteration the algorithm will try to diminish the value of loss function. Both AdaBoost and GBDT are consecutive ensemble models. In contrast, random forest, by concept, is a parallel ensemble model, and its purpose is to build numerous independent weak classifier and average their results. They all belong to ensemble learning methods, which help improving machine learning results by combining several base models to produce an optimal predictive model. Generally ensemble learnings tend to have better performances and are less likely to overfit.

The machine learning frame that we use is scikit-learn (Pedregosa et al. (2011); formerly scikits.learn). All models require hyperparameters setting which is independent of the dataset. Hyperparameters optimization is mostly implemented empirically. However, Grid Search is a practical approach without much experience. In Table 1 we present our best hyperparameters derived with sklearn.model_selection.GridSearchCV below for each model (other parameters not mentioned are left as default).

Among the spectroscopically identified sources, we randomly select 3000 quasars and 3000 stars to train our models. A sample of 600 quasars and 600 stars (an

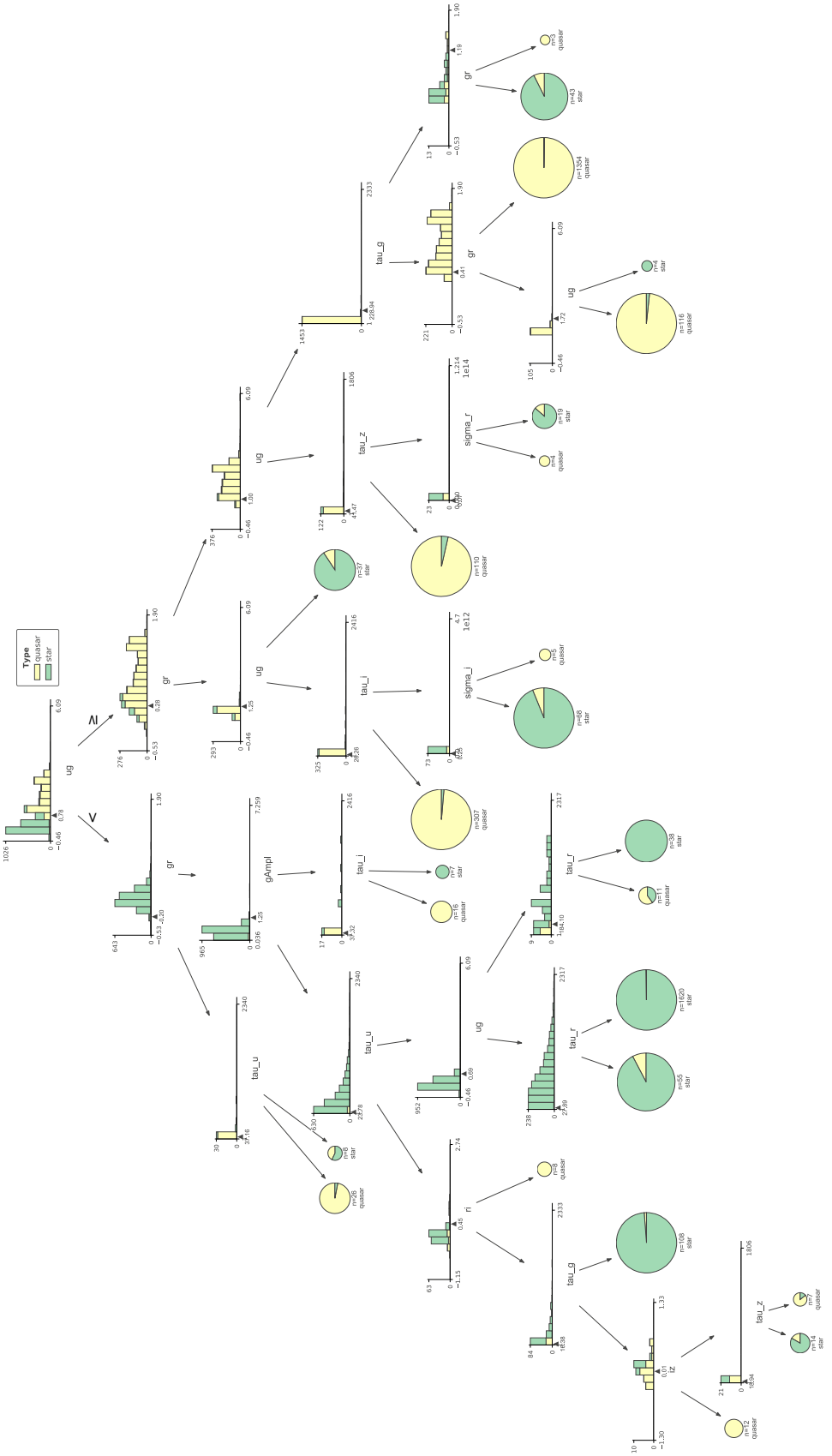

Fig. 1 The tree-like structure of an example decision tree in this work (pruned for better visualization). Visualization is done with python package **dtreeviz**.

Table 1 Hyperparameters Setting for All Models

Model	Hyperparameter	Value
RFC	n_estimators	500
	oob_score	*True*
AdaBoost	n_estimators	500
	learning_rate	0.08
	algorithm	*SAMME*
	max_depth	6
	min_samples_split	20
	min_samples_leaf	5
GBDT	n_estimators	500
	learning_rate	0.1

1:1 sample), mutually different from the training sample, is then built to test the trained classifiers. Two indices, namely precision and recall (see Eqs. (2)), are used to evaluate the performance of a trained model.

$$\begin{aligned} Precision_{star} &= \frac{\#\ of\ predicted\ true\ star}{\#\ of\ predicted\ star} \\ Recall_{star} &= \frac{\#\ of\ predicted\ true\ star}{\#\ of\ confirmed\ star} \\ Precision_{quasar} &= \frac{\#\ of\ predicted\ true\ quasar}{\#\ of\ predicted\ quasar} \\ Recall_{quasar} &= \frac{\#\ of\ predicted\ true\ quasar}{\#\ of\ confirmed\ quasar} \end{aligned} \quad (2)$$

where precision stands for the fraction of that a certain type of classifications is true, and recall the completeness of correct classification for a given class of objects. We repeat above random selections of training and test samples 100 times, and present the averaged performance in Section 4.1. Note we adopt a ratio of 1:1 of stars versus quasars for both the training and test datasets. While this is a common approach in machine learning studies, we discuss in Section 4.4 the effects of imbalanced datasets.

4 RESULTS AND DISCUSSION

4.1 The high performance of the machine learning algorithms

In Table 2, we present the performance of the three classifiers evaluated with the test samples. The values in the table are obtained through averaging 100 trials (hereafter the same), and the errors of the mean values are also presented. We find similarly high performances obtained with all three machines, showing each of them is competent enough for such a study. In Figure 2, we present the receiver operating characteristic (ROC) curve yielded by the RFC classifier, created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. The ROC curve appears ideal for a machine learning task, indicating the distributions of the two types of objects are well separated.

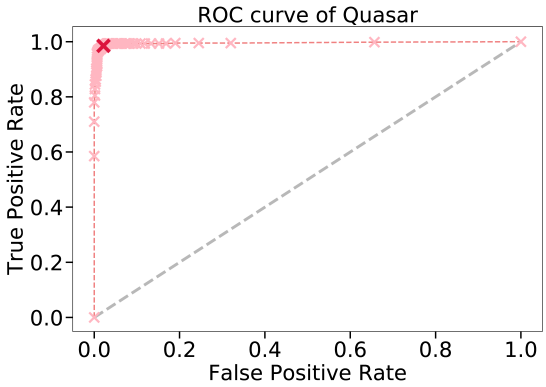

Fig. 2 ROC curve of quasar (i.e., "true sample" here means quasar) given by the RFC classifier. The red cross in the plot is the best threshold 0.42, the nearest point to (0.0,1.0).

We also find that using the variability features alone can yield better performances comparing with using only color features. This clearly demonstrates the high efficiency of selecting quasars with time domain observations. Putting variability and color features together would further enhance the selection accuracy, with $\sim$ 99.0% precision and recall achieved for all three machines.

Next, we briefly compare our results with previous representative works on quasar selection out of SDSS stripe 82 variable sources.

MacLeod et al. (2011) selected 10 024 SDSS Stripe 82 variable sources with $i < 19.0$. Among them 1490 ($\sim$ 15%) are spectroscopically confirmed quasars, and the rest were considered as non-quasars. They found that simple cuts in DRW parameters (such as $\tau > 100$ days), aided with the color selection, could achieved a precision of 93.8% (named as efficiency in the paper) and recall of 98.0% (named as completeness) in selecting quasars [2]. After corrected for the imbalance (see Sect.4.4), their precision and recall are 99.3% and 93% respectively. It can be seen that the analytical approach of MacLeod et al. (2011) utilized rather strict thresholds which could achieve high precision, but suffer considerable incompleteness.

Takata et al. (2018) used SDSS Stripe 82 variable sources catalog to train a supportive vector machine (SVM). They constructed a dataset which consists of 7714 confirmed quasars and 2141 stars. They used various sets of variability features to train the SVM. Their test dataset contains 1000 quasars and 1000 stars, and the rest of the sources were used to train the machine. Using DRW parameters measured with JAVELIN, they obtained averaged precision and recall of 93.8% and 98.6% [3]. While their recall is similar to our results, their precision is

[2] Note the performance was estimated with the same dataset that they used to define the thresholds, thus it could have been over-estimated.

[3] Private communications with the authors show that there are mistakes with their definitions of precision and recall in their original

Table 2 The performance of three machine learning algorithms. The mean values (of precision and recall) and the corresponding errors of the mean values are derived through averaging results from 100 trials. In the "all features" model both color and variability features (see Sect. 3.1) are utilized.

feature mode		RFC	AdaBoost	GBDT
all features	P_{quasar}	$99.00\% \pm 0.04\%$	$99.05\% \pm 0.04\%$	$98.90\% \pm 0.04\%$
	R_{quasar}	$98.84\% \pm 0.04\%$	$99.02\% \pm 0.04\%$	$98.85\% \pm 0.04\%$
	P_{star}	$98.84\% \pm 0.04\%$	$99.02\% \pm 0.04\%$	$98.85\% \pm 0.04\%$
	R_{star}	$99.00\% \pm 0.04\%$	$99.05\% \pm 0.04\%$	$98.90\% \pm 0.04\%$
color features	P_{quasar}	$97.16\% \pm 0.06\%$	$97.20\% \pm 0.06\%$	$97.23\% \pm 0.06\%$
	R_{quasar}	$96.64\% \pm 0.07\%$	$96.56\% \pm 0.07\%$	$96.43\% \pm 0.07\%$
	P_{star}	$96.66\% \pm 0.06\%$	$96.59\% \pm 0.06\%$	$96.46\% \pm 0.07\%$
	R_{star}	$97.18\% \pm 0.06\%$	$97.22\% \pm 0.07\%$	$97.24\% \pm 0.06\%$
variability features	P_{quasar}	$98.56\% \pm 0.04\%$	$98.43\% \pm 0.04\%$	$98.31\% \pm 0.05\%$
	R_{quasar}	$97.50\% \pm 0.06\%$	$97.81\% \pm 0.05\%$	$97.59\% \pm 0.06\%$
	P_{star}	$97.53\% \pm 0.06\%$	$97.82\% \pm 0.05\%$	$97.61\% \pm 0.06\%$
	R_{star}	$98.57\% \pm 0.04\%$	$98.44\% \pm 0.04\%$	$98.32\% \pm 0.05\%$

considerably lower. This could be partly due to the fact that they used smaller sample of 1141 stars to train the classifier. Also, SVM is a weak classifier, like a single decision tree in the random forest model. For comparison, the precision and recall of quasar given by a single decision tree are $\sim 98.0\%$ and $\sim 97.0\%$. Meanwhile, their precision is considerably lower than their recall, mainly because of the imbalanced test sets.

Graham et al. (2014) used SWV, DRW and SF at the same time to describe the variability features. Using variability alone, they obtained 96.5% recall (named as completeness in the literature) and 95.0% precision (named as purity in the literature) for their RFC. With the aid of color features, they obtained 99.3% recall and 99.0% precision. Using DRW parameters alone, we achieve similar recall and precision, confirming the results of Graham et al. (2014).

4.2 The completeness of the spectroscopically confirmed quasar sample in Stripe 82

We then apply our trained models to select potential quasars out of the unlabeled sources in the Stripe 82 Variable Source Catalog. With RFC, we classify 1105 of them as predicted quasars, and 47 501 as stars (As mentioned in Sect. 3.1, the DRW fitting failed for 110 unlabelled sources, which are most likely to be stars, thus having no influence on further discussion on completeness). Similar numbers are obtained using AdaBoost and GBDT. We plot the i-band magnitude distributions of the RFC predicted quasars and stars in Figure 3, together with those of the confirmed ones. While confirmed quasars and stars have significantly different magnitude distributions, it is interesting to note that the i-band magnitude distributions of predicted and spectroscopically confirmed quasars are similar, and so do those of stars.

Recalls, i.e., $\frac{\#\ of\ true\ predicted\ *}{\#\ of\ real\ *}$ (*: quasar / star), do not change with the absolute values of the numbers of quasar and star in the samples, whereas precisions, i.e., $\frac{\#\ of\ true\ predicted\ *}{\#\ of\ predicted\ *}$, may be affected greatly due to the imbalance, as mentioned in Section 4.4. Assuming recalls of the unlabeled sources are the same as those in Table 2, we can estimate the numbers of real sources with Equations (3), which are derived from the definitions of recalls in Equations (2). After correcting the incompleteness and contaminations due to misclassifications, we finally expect there are $\sim$ 633 real quasars among the 48 716 unlabeled sources, $\sim 633 \times (1 - 0.9984) \sim 7$ of them could have been misclassified as stars, and the sample of the 1105 predicted quasars has an precision of $\sim \frac{633-7}{1105} \times 100\% \sim 57\%$. This suggests the spectroscopically confirmed sample of quasars among the variable sources has a completeness of $\sim \frac{8330}{8330+633} \times 100\% \sim 93.0\%$, similar to the estimated completeness of SDSS quasars from small spectroscopically complete samples (>90%, Richards et al. 2002; Ivezić et al. 2002; Vanden Berk et al. 2005; Peters et al. 2015).

$$\begin{aligned}
&(1 - Recall_{star}) \times \#\ of\ real\ star = \\
&\#\ of\ predicted\ quasar - Recall_{quasar} \times \#\ of\ real\ quasar \\
&(1 - Recall_{quasar}) \times \#\ of\ real\ quasar = \\
&\#\ of\ predicted\ star - Recall_{star} \times \#\ of\ real\ star
\end{aligned} \tag{3}$$

We see from Figure 3 that, among the variables sources, stars significantly outnumber quasars at brighter magnitudes; and at the faint end, quasars are the dominant population. We thus expect that at brighter magnitudes, the predicted quasar sample suffers stronger contaminations from misclassified stars. This effect could be corrected through repeating the calculations described in the above paragraph, but at different limiting magnitudes.

In Figure 4 we plot the recalls of quasars and stars we measured with test samples, as a function of limiting i band

published paper, and an erratum is to be submitted. The values that we quoted here are corrected ones given by the authors.

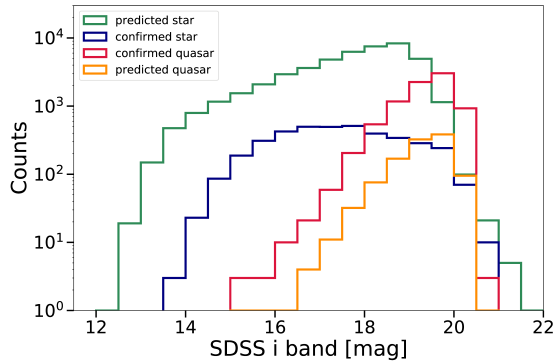

Fig. 3 SDSS i-band magnitude distributions of spectroscopically confirmed and RFC predicted quasars and stars. Predicted sources are those without spectroscopical identifications, but classified by our random forest model.

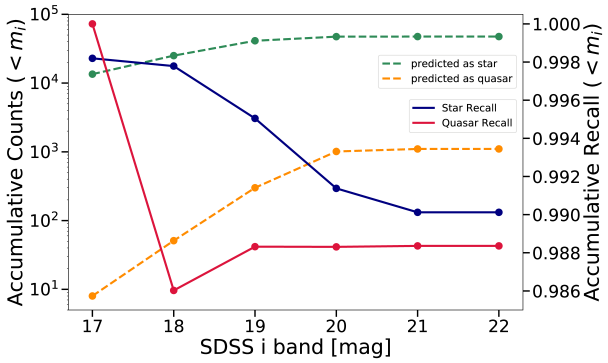

Fig. 4 The RFC performance (accumulative recall, averaged through 100 testing runs), and the numbers of predicted quasars/stars (out of the unlabeled sources), as a function of limiting magnitude m_i.

magnitude. Utilizing the measured recall, and the number of predicted quasars and stars, we calculate the corrected completeness (as a function of limiting magnitude, i.e. $< m_i$) in Figure 5. This indicates that SDSS quasar sample is highly complete ($\sim$ 95%) at $i < 19$, which is also consistent with the estimated completeness based on small spectroscopically complete samples (Vanden Berk et al. 2005, 94.9% for $i < 19.1$; and Peters et al. 2015, 94.7% for $i < 19.1$).

The "original completeness", calculated simply using the numbers of predicted quasars from our RFC classifier, is also plotted in Figure 5 for comparison. Note that this completeness is significantly contaminated by stars which have been misclassified as quasars, and such contamination is stronger at brighter magnitudes as there are relatively more bright stars than quasars in the variable source catalog (see Fig. 3). This effect could explain the even smaller "original completeness" at brighter magnitudes in Figure 5.

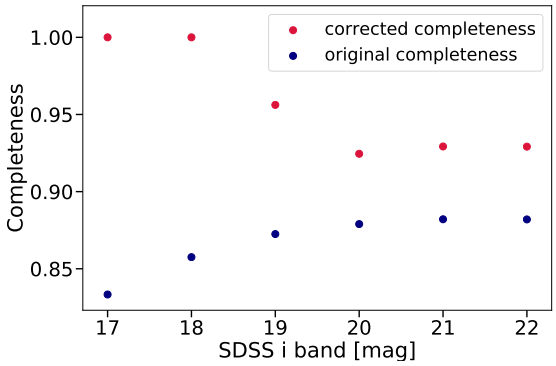

Fig. 5 The estimated completeness of spectroscopically confirmed quasars in the SDSS Stripe 82 variable source catalog, as a function of limiting magnitude m_i. The "original completeness" is calculated simply using the numbers of predicted quasars from our RFC classifier, i.e., (spectroscopically confirmed)/(confirmed + predicted).

4.3 Feature Importance

Not all of the input features to the classifiers are equally useful in distinguishing quasars from stars. It is helpful to examine the relative importance of various features, particularly considering that the available features are practically often limited by observational resources.

A decision tree or random forest can generate features rankings by calculating so called "gini importance" or "mean decrease impurity" (Breiman et al. 1984), which calculates each feature importance as the sum over the number of splits (across all tress in a random forest) that include the feature, proportionally to the number of samples it splits. In other words, features that can separate a larger set of samples into two pure enough subsets have larger feature importances. We run 100 trials to get average scores of every features. We present the results in Figure 6, where significant diversity in the feature scores is seen.

We then add feature one at a time into the feature set in the order of their importance scores, to train and test the RFC model. The output $precision$ and $recall$ are plotted in Figure 7, where we clearly see that the performance of the classifier is dominantly driven by the first few features.

In Figure 8 we also plot the measured $precisions$ and $recalls$ by using variation features in one band only. We see that g and r bands have similarly good performance, likely because SDSS quasars have the best photometry in g and r. Meanwhile, the performance in z band is the worst, which could be attributed to its much worse photometry and the fact that quasar variations are much weaker at longer wavelength. Note that the performance using only g or r band variation features alone is already comparable to that using all SDSS colors.

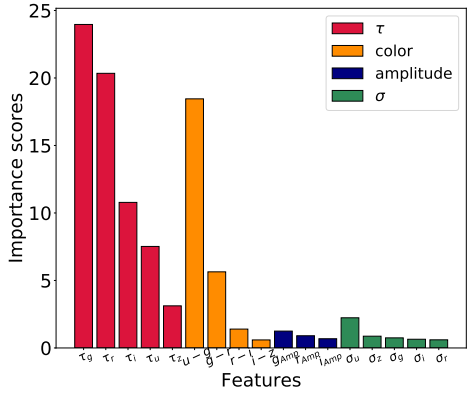

Fig. 6 Ranking of feature importance (gini importance) given by RFC. Red (green) bars represent DRW parameter τ (σ) in five bands, yellow bars the maximum variation amplitudes in gri, and blue bars the color features.

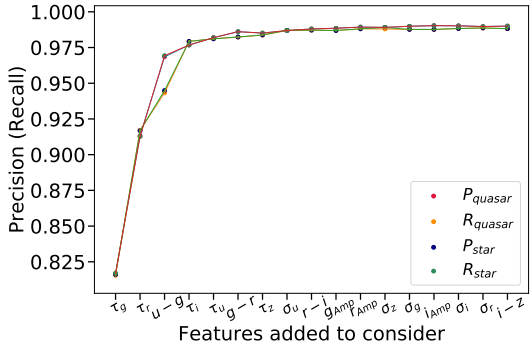

Fig. 7 *Precisions* and *recalls* of different input feature sets. Features are added one at a time by the ranking of importance score (from left to right). Only τ_g is used in the first point, and all features are included in the last point. We notice a sharp rise in the first three points, for they represent the top three most importance features.

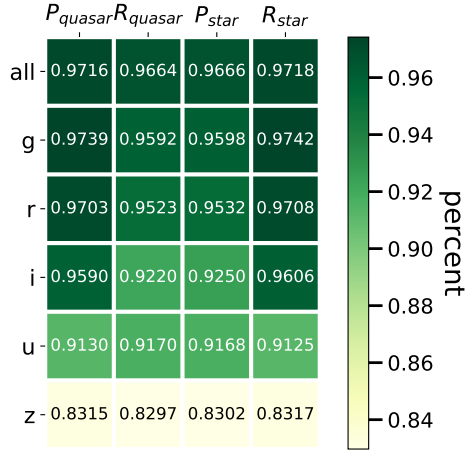

Fig. 8 The achieved performance of RFC of using variation features from one band only, e.g., a block marked by g represents performances that only variation features of *g-band* are used, which are τ_g, g_{Amp} and σ_g.

4.4 Effects of Imbalanced Samples

We previously present and discuss the efficiencies of the machine learning models using training and test samples with 1:1 ratio of quasars and stars. However actual datasets are often heavily imbalanced. For instance, among the SDSS Stripe 82 variable sources we used in this work, there are 8330 spectroscopically confirmed quasars, but only 3966 stars. More significantly, among the 48 716 unlabeled sources, we only expect $\sim$ 630 quasars (see Sect. 4.2). Below we discuss the effects of sample imbalance in both the test sample and the training sample.

The effect of imbalance in the test sample (or the to-be-classified sample) is straightforward. Comparing with an 1:1 sample, an imbalanced test sample would in principle yield identical *recall* but different *precision* for each type of the objects. This is because, for instance, whether a quasar in the test sample could be correctly classified (*recall*) depends on the observed features of the quasar and how the classifier was trained, but has nothing to do with other sources in the test sample. In contrast, the *precision* of the quasar classifications does depend on the number of stars which have been mis-classified as quasars, thus the number of stars in the sample.

Let η be the ratio of stars to quasars in the test sample. The expected precisions of quasars and stars from an imbalanced test sample can be expressed as:

$$Precision_{quasar} = \frac{Recall_{quasar}}{Recall_{quasar} + \eta \times (1 - Recall_{star})}$$
$$Precision_{star} = \frac{Recall_{star}}{Recall_{star} + 1/\eta \times (1 - Recall_{quasar})} \quad (4)$$

We can clearly see from the above equation that, a test sample with stars more than quasars ($\eta > 1$) would yield lower quasar precision (and higher star precision) comparing with the 1:1 sample (see also Sect. 4.2).

The effect of imbalanced training sample is more complicated and there is no simple analytical equation. In principle, if there are more stars than quasars in the training set, then the machine learning model will likely learn more information about the stars. In this way, a star will be less likely misclassified as a quasar. But apparently, the shortage of this approach is that a quasar will be more likely misclassified as a star. This compromise shows the well-known trade between precision and recall. This imbalance is a common issue in the machine learning field, and many algorithms and methods have been brought forward to deal with it (e.g. He & Garcia 2009; Chawla et al. 2004). Generally, utilizing special sampling methods (e.g. resampling, over-sampling, under sampling) and

Table 3 The performance (precision and recall, similar to Table 2) of three trained RFC models using one-year, two-year and ten-year long light curves.

feature mode		one-year	two-year	ten-year
all features	P_{quasar}	98.18% ± 0.05 %	98.29% ± 0.05 %	98.91% ± 0.04 %
	R_{quasar}	98.09% ± 0.05 %	98.41% ± 0.05 %	98.89% ± 0.04 %
	P_{star}	98.10% ± 0.05 %	98.41% ± 0.05 %	98.89% ± 0.04 %
	R_{star}	98.18% ± 0.05 %	98.29% ± 0.05 %	98.90% ± 0.04 %
variability features	P_{quasar}	95.73% ± 0.07%	97.13% ± 0.07 %	98.54% ± 0.04 %
	R_{quasar}	94.87% ± 0.09%	96.16% ± 0.08 %	97.74% ± 0.05 %
	P_{star}	94.92% ± 0.08%	96.20% ± 0.08 %	97.76% ± 0.05 %
	R_{star}	95.73% ± 0.08%	97.15% ± 0.07 %	98.55% ± 0.04 %
color features	P_{quasar}	—	—	97.08% ± 0.06%
	R_{quasar}	—	—	96.68% ± 0.08%
	P_{star}	—	—	96.70% ± 0.07%
	R_{star}	—	—	97.09% ± 0.06%

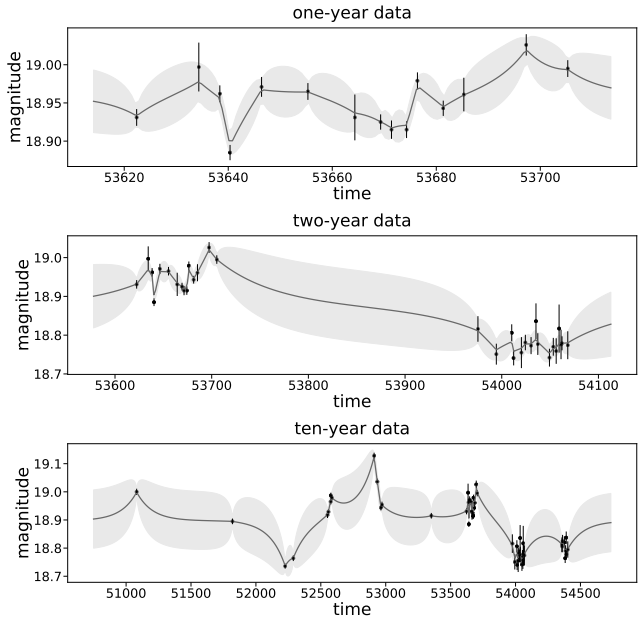

Fig. 9 From top to bottom: full ten-year, two-year (2005–2006), and one-year (2005) SDSS g band light curves of an example quasar, and the corresponding best-fitted DRW models.

adjusting loss function (e.g. cost-sensitive) are the most common ways to deal with the problem.

4.5 Dependence on the length of light curves

In future surveys, pre-selection of quasar candidates could be required when only one or two semesters of time domain observations are available. Below we explore whether the performance of the variability-based quasar selection is sensitive to the length of the light curves utilized to derived the variability parameters. This is realized through feeding one-year and two-year data extracted from the Stripe 82 light curves to the classifiers. The one-year data are specified as data collected in 2005, and two-year as data collected in 2005 and 2006. The typical number of epochs in one-year and two-year data are 15 and 30 respectively. Example quasar light curves are given in Figure 9.

Following the procedures described in Section 3.2, we train three RFC models using ten-year, two-year and one-year datasets respectively. For all models we use samples consisted of randomly selected 2800 quasars plus 2800 stars[4] to train, and 600 quasars plus 600 stars to test. With 100 trials we present the derived confusion matrixes with mean precision and recall in Table 3.

We find that when using shorter light curves, the derived test scores are lower, but only slightly. The lower scores are mainly because shorter light curves yielded smaller τ for quasars (see Fig. 10), making them harder to be distinguished from stars. Nevertheless, the performance of quasar selection by variability alone using two-year light curves is similar to that using color features alone. The one-year datasets yield only slightly worse scores, demonstrating that selecting quasars by variability is still efficient even when only one semester time domain observations are available (with $\sim$ 13 epochs for stars and $\sim$ 18 epochs for quasars).

Finally, we investigate the dependence of the performance on the number of epochs obtained within one observing semester. We select sources that have at least 15 epochs in 2005. 6836 quasars and 1624 stars are selected. We then randomly select 5, 9, 15 epochs in 2005 from each source to fit with a DRW. Using variability features alone, we use 1300 quasars and stars to train and 300 to test. The results averaged after 100 trials are listed in Table 4. The results of "all epochs" are comparable to those of "one-year variability features" in Table 3, but with higher R_{quasar}, P_{star} and lower P_{quasar}, R_{star}. This is because though the new subsamples (with at least 15 epochs from each source) are smaller than the "one-year" samples used in 3 that we can use only 1300 quasars plus 1300 stars to train (instead of 2800:2800), the new subsamples have on

[4] Here we use 2800:2800 sources (instead of 3000:3000) to train the classifiers as for some stellar sources the DRW fitting fails for one-year or two-year datasets.

Table 4 The performance of RFC using data with different length (variability features alone). From left to right are: 5 epochs, 9 epochs, 15 epochs (all random selected) and all epochs in 2005.

	5 epochs	**9 epochs**	**15 epochs**	**all epochs**
P_{quasar}	90.55% ± 0.15 %	92.54% ± 0.14 %	93.74% ± 0.12 %	94.51% ± 0.13 %
R_{quasar}	92.10% ± 0.17 %	93.71% ± 0.14 %	94.91% ± 0.13 %	95.54% ± 0.13 %
P_{star}	91.98% ± 0.16 %	93.65% ± 0.14 %	94.86% ± 0.13 %	95.51% ± 0.13 %
R_{star}	90.36% ± 0.22 %	92.43% ± 0.15 %	93.65% ± 0.13 %	94.43% ± 0.13 %

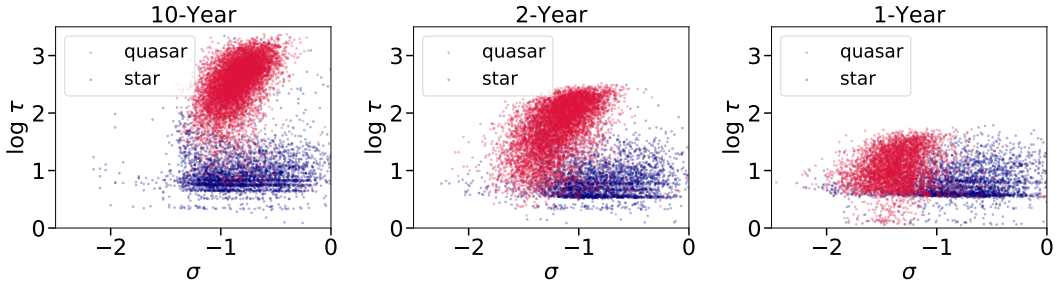

Fig. 10 τ vs σ for different datasets. Left to right: DRW parameters derived from ten-year, two-year and one-year g band light curves, respectively. A similar version of the left panel could be seen in MacLeod et al. (2011).

average more epochs from each source from those "one-year" samples used in Table 3. Clearly from Table 4 we can see that the performance decreases with decreasing number of epochs. Note using even "five epochs" within one semester could still reach considerably high precision and recall ($\sim 90\%$), further demonstrating the efficiency of variation-based quasar selection.

5 CONCLUSIONS

In this work, we extensively study variability-based quasar selection through training and testing three data-driven classifiers (random forest, AdaBoost, GBDT) with the 10-year long multi-epoch optical photometric data in SDSS Stripe 82. We fit the SDSS Stripe 82 light curves of spectroscopically confirmed quasars and stars with the DRW process using JAVELIN. The main results of this work include:

1. Trained with the variability features alone, all three models can select quasars with similarly and remarkably high precision and completeness ($\sim$ 98.5% and 97.5%, trained and tested with 1:1 samples of quasars and stars), even better than using SDSS colors alone ($\sim$ 97.2% and 96.5%).

2. Combining both variability and color features, we achieve precision and completeness both $\sim$ 99.0%, consistent with previous similar studies.

3. Using the trained models, we classify the unlabelled variable sources in Stripe 82, and estimate the completeness of the spectroscopically identified quasar sample in Stripe 82 variable source catalog to be $\sim$ 95% (for $m_i <$ 19.0).

4. We present the relative importance of the observational features utilized to classify quasars. The top three most important features are τ_g, τ_r, and $u - g$.

5. We show that variability-based quasar selection could still be highly efficient even when only one- or two-year time domain observations are available.

We also discuss the effects of imbalanced samples used to train and test the classifiers.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant Nos. 11421303 and 11890693), the National Basic Research Program of China (973 program, Grant No. 2015CB857005) and CAS Frontier Science Key Research Program (QYZDJ-SSW-SLH006).

Appendix A: PECULIAR FITTED DRW PARAMETERS

In Figure A.1, we show an example stellar source whose fitted σ is extremely large. From the light curves, we presume that these kind of sources could be cataclysmic variable stars, dwarf novae, etc. Due to the strong variation at very short time scales, the fitted DRW parameters could be abnormal, and because of the same reason, they are easy to distinguish.

References

Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42

Andrae, R., Kim, D. W., & Bailer-Jones, C. A. L. 2013, A&A, 554, A137

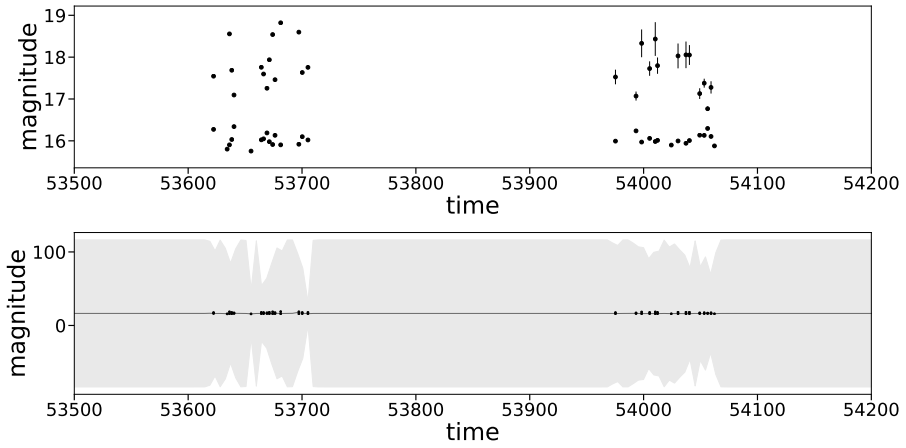

Fig. A.1 Top: An example SDSS g band light curve of a star with extremely large fitted σ. Bottom: best-fitted DRW model from JAVELIN. Only observations between 53500-54200 are shown in both panels.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. 1984, Classification and Regression Trees, The Wadsworth and Brooks-Cole statistics-probability series (Taylor & Francis)

Butler, N. R., & Bloom, J. S. 2011, AJ, 141, 93

Chawla, N. V., Japkowicz, N., & Kotcz, A. 2004, ACM SIGKDD Explorations Newsletter, 6, 1

Choi, Y., Gibson, R. R., Becker, A. C., et al. 2014, ApJ, 782, 37

Donley, J. L., Koekemoer, A. M., Brusa, M., et al. 2012, ApJ, 748, 142

Freund, Y., & Schapire, R. E. 1995, A Desicion-theoretic Generalization of On-line Learning and an Application to Boosting, in Computational Learning Theory, ed. P. Vitányi (Berlin, Heidelberg: Springer Berlin Heidelberg), 23

Friedman, J. H. 2002, Computational Statistics & Data Analysis, 38, 367 ,

Graham, M. J., Djorgovski, S. G., Drake, A. J., et al. 2014, MNRAS, 439, 703

Guo, H., Wang, J., Cai, Z., & Sun, M. 2017, ApJ, 847, 132

He, H., & Garcia, E. A. 2009, IEEE Transactions on Knowledge and Data Engineering, 21, 1263

Hernitschek, N., Schlafly, E. F., Sesar, B., et al. 2016, ApJ, 817, 73

Ivezić, Ž., Menou, K., Knapp, G. R., et al. 2002, AJ, 124, 2364

Ivezić, Ž., Smith, J. A., Miknaitis, G., et al. 2007, AJ, 134, 973

Kasliwal, V. P., Vogeley, M. S., & Richards, G. T. 2017, MNRAS, 470, 3027

Kelly, B. C., Bechtold, J., & Siemiginowska, A. 2009, ApJ, 698, 895

Kelly, B. C., Becker, A. C., Sobolewska, M., Siemiginowska, A., & Uttley, P. 2014, ApJ, 788, 33

Kelly, B. C., Sobolewska, M., & Siemiginowska, A. 2011, ApJ, 730, 52

Kim, D.-W., Protopapas, P., Byun, Y.-I., et al. 2011, ApJ, 735, 68

Kozłowski, S., Kochanek, C. S., Udalski, A., et al. 2010, ApJ, 708, 927

Lacy, M., Petric, A. O., Sajina, A., et al. 2007, AJ, 133, 186

MacLeod, C. L., Ivezić, Ž., Kochanek, C. S., et al. 2010, ApJ, 721, 1014

MacLeod, C. L., Brooks, K., Ivezić, Ž., et al. 2011, ApJ, 728, 26

Mason, L., Baxter, J., Bartlett, P., & Frean, M. 1999, Boosting Algorithms as Gradient Descent in Function Space

Mushotzky, R. F., Edelson, R., Baumgartner, W., & Gand hi, P. 2011, ApJL, 743, L12

Pâris, I., Petitjean, P., Aubourg, É., et al. 2018, A&A, 613, A51

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Peters, C. M., Richards, G. T., Myers, A. D., et al. 2015, ApJ, 811, 95

Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945

Sánchez-Sáez, P., Lira, P., Cartier, R., et al. 2019, ApJS, 242, 10

Schmidt, K. B., Marshall, P. J., Rix, H.-W., et al. 2010, ApJ, 714, 1194

Schneider, D. P., Richards, G. T., Hall, P. B., et al. 2010, AJ, 139, 2360

Sesar, B., Ivezić, Ž., Lupton, R. H., et al. 2007, AJ, 134, 2236

Simm, T., Salvato, M., Saglia, R., et al. 2016, A&A, 585, A129

Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163

Takata, T., Mukuta, Y., & Mizumoto, Y. 2018, ApJ, 869, 178

Vanden Berk, D. E., Schneider, D. P., Richards, G. T., et al. 2005, AJ, 129, 2047

Yao, S., Wu, X.-B., Ai, Y. L., et al. 2019, ApJS, 240, 6

Zhu, S. F., & Xue, Y. Q. 2016, ApJ, 825, 56

Zu, Y., Kochanek, C. S., Kozłowski, S., & Udalski, A. 2013, ApJ, 765, 106

Zu, Y., Kochanek, C. S., & Peterson, B. M. 2011, ApJ, 735, 80