Machine Learning for Quasar Candidate Selection

Yanxia Zhang, Jingyi Zhang and Yongheng Zhao

Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, 20A Datun Road, Chaoyang District, 100012, Beijing, P.R.China; zyx@bao.ac.cn

Abstract. With the big data era of astronomy coming, machine learning becomes more popular in various astronomical aspects, for example, classification of celestial objects, physical parameter measurement, rare object detection and so on. Quasar candidate selection belongs to the classification problem. Based on a number of present photometric and spectroscopic sky survey projects, preselecting quasar candidates from galaxies and stars is a hot issue. It is easy to separate quasars and stars from galaxies by morphology while it is difficult to distinguish quasars from stars due to their similar morphology. Nevertheless, researchers have figured out various approaches on quasar candidate selection. We will introduce them in detail.

1. Introduction

With the development of large sky survey projects, such as SDSS, LAMOST, 2MASS and WISE, the number of known quasars has increased quickly in recent years and added up to more than 500,000. This achievement mainly depends on the improvement of observational technologies, the availability of multiwavelength data, the application of machine learning in targeting quasar candidates. The past aim of quasar discovery is to find more quasars while the recent aim is to find unusual quasars, for example, high redshift quasars, extremely luminous quasars. The acquisition of quasars are important to understand the physical mechanism, formation and evolution of quasars and the early history of the Universe.

2. Schemes to target quasar candidates

So far there have been various methods to pick out quasar candidates. The quasar candidates selected from single band are inclined to be some kind of quasars or quasars with some range of redshift. To overcome the disadvantage, the application of multiwavelength data is a good solution, for example, optical information combined with infrared information. By the distribution of known galaxies, stars and quasars in the color-color spaces or color-magnitude spaces, the criterion to separate these three kinds of objects may be obtained, then it is applied on the unknown objects and select out quasar candidates, as show in Figure 1. But these criterions based on color-color cut or color-magnitude cut only consider the information in a two dimensional space, not using all features provided by the samples. For this reason, machine learning algorithms are adopted. The schemes are shown in Figure 2. The known sample may include

galaxies, stars and quasars (left panel of Figure 2), or stars and quasars after ruling out galaxies by color-color cut (right panel of Figure 2). Such algorithms with the successful applications contain Support Vector Machine (SVM; (Peng et al. 2012)), Kernel Density Estimation (KDE; (Richards et al. 2004)), Neural Network (NN; (Yèche et al. 2010)), Extreme Deconvolution (XD; (Bovy et al. 2011)), random forest ((Schindler et al. 2017)) and so on. In this scheme, only one machine learning algorithm is implemented. Each algorithm has its inclination. So the machine learning algorithm committee may replace single machine learning algorithm, as indicated in Figure 3. Nevertheless the effectiveness of selecting quasar candidates may be improved while its efficiency will decrease due to high-cost computation.

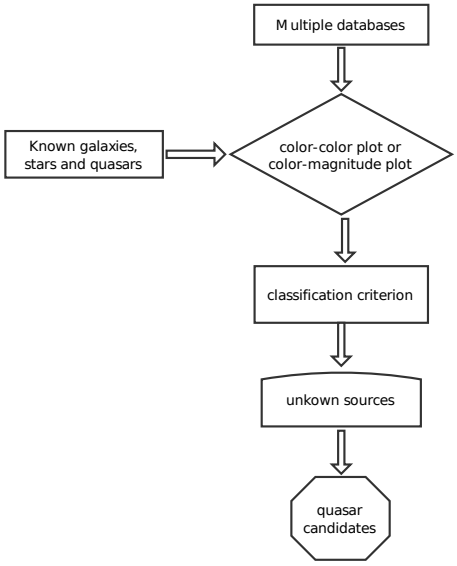

Figure 1. The scheme of color-color cut or color-magnitude cut

Some researchers focus on the study of high redshift quasars. Nevertheless the number of known high redshift quasars is too small. In order to increase this sample, different approaches are figured out, for example, color-color cut, infrared-based methods. Through above schemes (Figures 1-3), quasar candidates are available. Based on the known quasar sample, we may create a multiple classifier to divide the quasar candidates into low redshift quasars, medium redshift quasars and high redshift quasars, as shown in the left panel of Figure 4, or build a regressor to predict the redshifts of the quasar candidates as described in the right panel of Figure 4. The regression algorithms fit for photometric redshift estimation of quasars include k Nearest Neighbors (kNN; (Zhang et al. 2013)), Skew-t ((Schindler et al. 2017)), random forest ((Schindler et al. 2017)), Support Vector Machine (SVM; (Schindler et al. 2017)), combined kNN and SVM ((Han et al. 2016)), and so on.

No matter for classification issues or for regression issues, appropriate choice of algorithms is a must. Drawing lessons from the past examples, we select the best algorithm, or improve the old methods, even develop new methods. In reality, we should consider both effectiveness and efficiency. The more data available, the more features can be used. Facing lots of features, feature selection or feature extraction is neces-

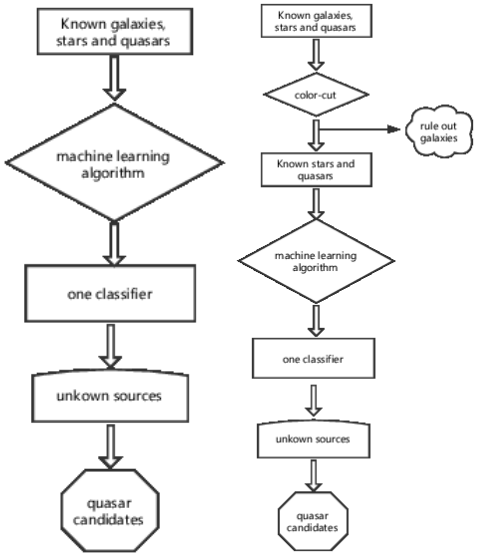

Figure 2. The scheme of machine learning algorithm

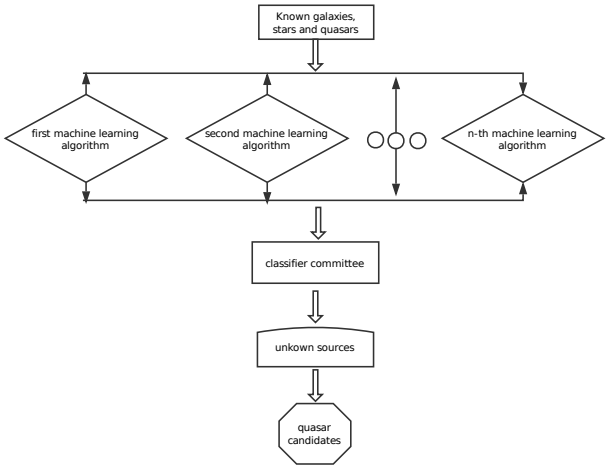

Figure 3. The scheme of machine learning algorithm committee

sary ((Zheng & Zhang 2008)). According to the characteristics of data and adopted algorithms, we carefully perform feature selection or feature extraction.

3. Conclusion

We summarize the schemes to target quasar candidates. By means of these schemes, more and more quasars will be discovered with the continuous increase of various large sky survey data. Most works on this respect make use of single epoch data. The time series data from Gaia, Pan-STARRS and upcoming LSST will provide new opportunity for quasar discovery.

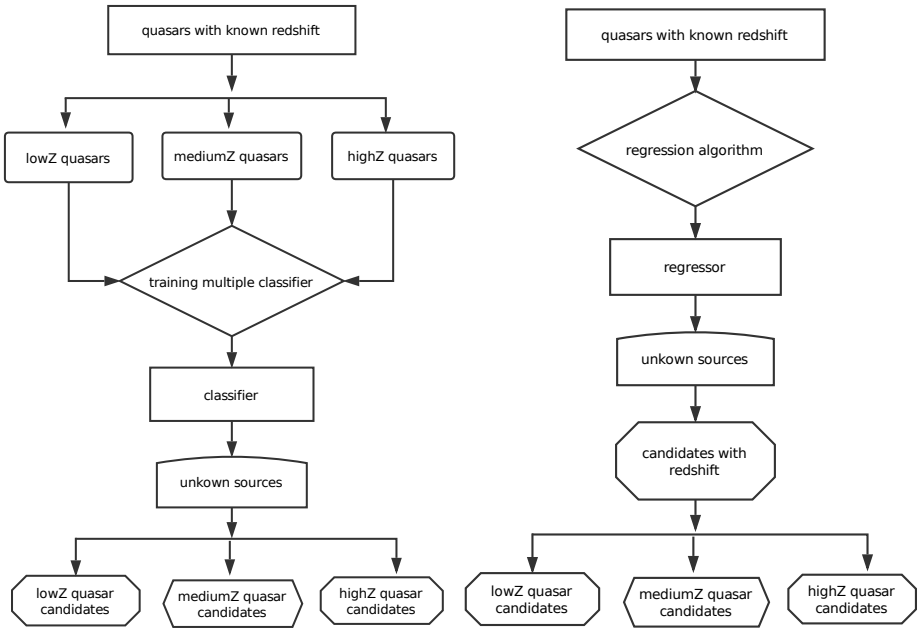

Figure 4. The scheme to discriminate different redshift quasars

Acknowledgments. This paper is funded by National Key Basic Research Program of China 2014CB845700, National Natural Science Foundation of China under grant Nos. 11873066 and U1731109.

References

Bovy, J., Hennawi, J. F., Hogg, D. W., Myers, A. D., Kirkpatrick, J. A., Schlegel, D. J., Ross, N. P., Sheldon, E. S., McGreer, I. D., Schneider, D. P., & Weaver, B. A. 2011, ApJ, 729, 141. 1011.6392

Han, B., Ding, H.-P., Zhang, Y.-X., & Zhao, Y.-H. 2016, Research in Astronomy and Astrophysics, 16, 74. 1601.01739

Peng, N., Zhang, Y., Zhao, Y., & Wu, X.-b. 2012, MNRAS, 425, 2599. 1204.6354

Richards, G. T., Nichol, R. C., Gray, A. G., Brunner, R. J., Lupton, R. H., Vanden Berk, D. E., Chong, S. S., Weinstein, M. A., Schneider, D. P., Anderson, S. F., Munn, J. A., Harris, H. C., Strauss, M. A., Fan, X., Gunn, J. E., Ivezić, Ž., York, D. G., Brinkmann, J., & Moore, A. W. 2004, ApJS, 155, 257. astro-ph/0408505

Schindler, J.-T., Fan, X., McGreer, I. D., Yang, Q., Wu, J., Jiang, L., & Green, R. 2017, ApJ, 851, 13. 1712.01205

Yèche, C., Petitjean, P., Rich, J., Aubourg, E., Busca, N., Hamilton, J.-C., Le Goff, J.-M., Paris, I., Peirani, S., Pichon, C., Rollinde, E., & Vargas-Magaña, M. 2010, A&A, 523, A14

Zhang, Y., Ma, H., Peng, N., Zhao, Y., & Wu, X.-b. 2013, AJ, 146, 22. 1305.5023

Zheng, H., & Zhang, Y. 2008, Advances in Space Research, 41, 1960. 0709.0138