Probing Galaxies through Quasar Absorption Lines
Proceedings IAU Colloquium No. 199, 2005
P. R. Williams, C. Shu, and B. Ménard, eds.

doi:10.1017/S1743921305003157

Preselecting AGN candidates from multi-wavelength data by ADTree

Yanxia Zhang[1], Hongwen Zheng[2], and Yongheng Zhao[1]

[1]National Astronomical Observatories, Chinese Academy of Sciences, P. R. China.
email: zyx@lamost.org, yzhao@lamost.org

[2]Institute of Mathematics and Physics, North China Electric Power University, Beijing, P. R. China

Abstract. With the information era in astronomy coming, this "data avalanche" may provide many answers to important problems in contemporary astrophysics. The most important problem is sifting through massive amounts of data to mine knowledge. In this paper, we positionally cross-identify multi-wavelength data from optical, near-infrared, and X-ray bands, and then employ Alternating Decision Trees (ADTree) to quickly and robustly separate AGN candidates to a high degree of accuracy. We emphasise the application of the method due to the development of large survey projects and the establishment of the virtual observatory, and conclude that the application of data mining algorithms in astronomy is of great importance to discover new knowledge impossible to obtain before, and promote the development of astronomy.

1. Introduction

Because of their large intrinsic luminosity, an active galactic nucleus (AGN) can be observed up to very high redshifts, and AGNs can therefore be used to probe the early epochs of the Universe when the formation of large structures began. So it is necessary to construct a complete sample of AGNs to study this issue. AGNs are known for their wide spectral energy distribution (SED) so that we preselect them from different bands through their remarkable radiation properties.

While AGNs were first noticed to be a class of objects worthy of detailed follow-up in other bands on the basis of their radio emission (e.g. Schmidt 1963), more recent work has focused on surveys at shorter wavelengths. Because of their low spatial density, most AGNs at high redshift (defined here to be $z > 4$) were discovered by large-area, shallow optical surveys. However, optical surveys are subject to bias against the discovery of quasars that are either intrinsically dim at these wavelengths or are obscured by dust like the type-2 AGNs that are believed to comprise most of the AGN background (Hasinger 2002). The other main selection technique is X-ray detection, as X-ray emission appears to be a universal characteristic of AGNs at all observed redshifts (Kaspi *et al.* 2000). In addition, AGN candidate selection methods employed by previous surveys also include colour selection, slit-less spectroscopy (SS) selection, and selection by infrared sources, by variability, or by zero proper motion. In order to construct highly complete samples, combined methods have recently been employed. For example, the Large Bright Quasar Survey (LBQS; Hewett, Foltz & Chaffee 1995) used both colour and SS selection.

Faced with a "data avalanche" in astronomy, we need automated methods to preselect AGN candidates. Zhang *et al.* (2002, 2003a, 2003b) explored automated classification methods, Learning Vector Quantisation (LVQ), Support Vector Machine (SVM), and these two approaches combined with principal component analysis (PCA) to preselect AGN candidates. Their results add up to high accuracy.

2. Methodology

In this paper, we introduce a new type of classification rule, the alternating decision tree (ADTree), which is a generalisation of decision trees, voted decision trees, and voted decision stumps. A general alternating tree defines a classification rule as follows. An instance defines a set of paths in the alternating tree. As in standard decision trees, when a path reaches a decision node it continues with the child which corresponds to the outcome of the decision associated with the node. However, when reaching a prediction node, the path continues with all of the children of the node. More precisely, the path splits into a set of paths, each of which corresponds to one of the children of the prediction node. We call the union of all the paths reached in this way for a given instance the "multi-path" associated with that instance. The sign of the sum of all the prediction nodes that are included in a multi-path is the classification which the tree associates with that instance. The formal definition of alternating trees (ADTrees) can be referred to Freund & Mason (1999).

3. Data

The ROSAT Bright Source Catalogue (BSC; Voges *et al.* 1999) contains positions, X-ray count rates, and spectral information of 18,806 X-ray sources with count rates greater than 0.05 counts s^{-1}, observed during the ROSAT All-Sky-Survey (RASS). Similarly, the ROSAT Faint Source Catalogue (FSC) includes 105,924 sources. The catalogue of quasars and active nuclei (Véron-Cetty & Véron 2000) contains 13,214 quasars, 462 BL Lac objects, and 4,428 active galaxies (of which 1,711 are Seyfert 1).

We positionally cross-identify the Véron 2000 catalogue with the ROSAT Bright Source Catalogue (RASS/BSC) and Faint Source Catalogue (RASS/FSC) X-ray sources, and then cross-identify the result with optical sources in the USNO A-2.0 catalogue. Similarly, using these sources to positionally cross-match 2MASS released data, we cross out the one-to-many sources and get 909 quasars, 135 BL Lacs, and 612 active galaxies. By the same method, we adopt stars from SIMBAD and galaxies from the Third Reference Catalogue of Bright Galaxies (RC3; de Vaucouleurs *et al.* 1991) to obtain 3,718 stars and 173 normal galaxies from the optical, X-ray, and infrared bands. The chosen attributes from different bands are $B-R$ (optical index), $B+2.5\log(CR)$, $lgCR$ (source count-rate in the broad energy band), $HR1$ (hardness ratio 1), $HR2$ (hardness ratio 2), ext (source extent), $extl$ (likelihood of source extent), $J-H$ (infrared index), $H-K$ (infrared index), and $J+2.5\log(CR)$.

4. Results

We randomly split the data into two parts: 66% for training, the remainder for testing. The classification result is shown in Table 1. Here AGNs represent quasars, BL Lacs, and active galaxies; non-AGNs stand for stars and normal galaxies. The accuracy of AGNs adds up to 97.1%, that of non-AGNs is 96.9%. The whole accuracy amounts to 97.0%.

5. Conclusions

A novel automated classification method, the alternating decision tree (ADTree) has been applied in this paper. When applied to multi-wavelength data, the accuracy is up to 97.0%. So we can use ADTree to train the whole data to get classification rules and the rules can then be used for predicting types of new data or preselecting AGN

https://doi.org/10.1017/S1743921305003157 Published online by Cambridge University Press

Table 1. The classification result

classified ↓ known →	AGNs	non-AGNs
AGNs	544	41
non-AGNs	16	1285
accuracy	97.1%	96.9%

candidates. For the objects in which astronomers are interested, they can choose a known sample to train ADTree and get classifiers to preselect source candidates. The classifiers of this type are relatively easy to interpret, compared to other methods such as neural networks. Moreover the time taken to build a model is only 2.11 seconds. With the quantity, quality, and complexity of data improving, ADTree shows its efficiency and effectiveness. Especially faced with large sky surveys, automated methods can not only reduce astronomer's efforts, but also improve the efficiency of astronomers and high-cost telescopes. With the successful application of data mining in astronomy and the establishment of the International Virtual Observatory, astronomers will be able to do astronomy more easily and more conveniently.

Acknowledgements

This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France. Simultaneously, this paper has also made use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. This paper is funded by National Natural Science Foundation of China under grant No. 10473013.

References

de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G., *et al.*, 1991, Third Reference Catalogue of Bright Galaxies (RC3), New York: Springer-Verlag

Freund, Y., Mason, L., 1999, Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia, p. 124-133.

Hasinger, G., 2002, preprint (astro-ph/0202430)

Hewett, P. C., Foltz, C. B., Chaffee, F. H., 1995, AJ, 109, 1498

Kaspi, S., *et al.*, 2000, AJ, 119, 2031

Schmidt, M., 1963, Nature, 197, 1040

Voges, W., Aschenbach, B., Boller, Th., *et al.*, 1999, A&A, 349, 389

Véron-Cetty, M. P., Véron, P., 2000, ESO Scientific Report 19

Zhang, Y., Cui, C., Zhao, Y., 2002, in: Starck, Jean-Luc, Murtagh, & Fionn (eds.), Proc. SPIE, 4847, 371

Zhang, Y., Zhao, Y., 2003a, PASP, 115, 1006

Zhang, Y., Zhao, Y., 2003b, ChJAA, 3, 183