Chin. J. Astron. Astrophys. Vol. 7 (2007), No. 2, 289–296
(*http://www.chjaa.org*)

Chinese Journal of
Astronomy and
Astrophysics

A Comparison of BBN, ADTree and MLP in separating Quasars from Large Survey Catalogues *

Yan-Xia Zhang and Yong-Heng Zhao

National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012; *zyx@bao.ac.cn*

Received 2006 June 1; accepted 2006 December 25

Abstract We compare the performance of Bayesian Belief Networks (BBN), Multilayer Perception (MLP) networks and Alternating Decision Trees (ADtree) on separating quasars from stars with the database from the 2MASS and FIRST survey catalogs. Having a training sample of sources of known object types, the classifiers are trained to separate quasars from stars. By the statistical properties of the sample, the features important for classification are selected. We compare the classification results with and without feature selection. Experiments show that the results with feature selection are better than those without feature selection. From the high accuracy found, it is concluded that these automated methods are robust and effective for classifying point sources. They may all be applied to large survey projects (e.g. selecting input catalogs) and for other astronomical issues, such as the parameter measurement of stars and the redshift estimation of galaxies and quasars.

Key words: astronomical databases: miscellaneous — catalogs — methods: data analysis — methods: statistical

1 INTRODUCTION

The rapid emergence of huge, uniform, multivariate databases from specialized survey projects and telescopes has led to the coming of the 'information age' in astronomy, just like the 'data avalanche' faced in other fields. Powerful database systems for collecting and managing data are in use in virtually all large and mid-range astronomical institutes. How to collect, save, organize, and mine the data efficiently and effectively is an important issue. Due to the large size of the databases, it is almost impossible to manually analyze the data for new knowledge. Therefore, automated extraction of useful knowledge from huge amounts of data is now widely recognized, and leads to a rapidly developing market of automated analysis and discovery tools. Data mining and knowledge discovery are techniques to identify valid, novel, potentially useful, and ultimately understandable patterns hidden in very large databases. Automated discovery tools can provide the potential and advantages to mine raw data and obtain the extracted high level information to the analyst or astronomers.

As in most statistical tasks, the diversity of tasks and techniques in data mining is broad. For example, Fayyad et al. (1996) divided data mining tasks into six flavors: (i) classification; (ii) regression; (iii) clustering; (iv) summarization; (v) dependency modelling (structural and quantitative); and (vi) change modelling (changes from previous or normative values). In astronomy, the automatic classification of objects from catalogues is a common issue encountered in many surveys. From a list of values of variables associated with a celestial object, it is desired to identify the object's type (e.g. star and galaxy). In this paper we apply three automated methods: Bayesian Belief Networks (BBN), Multilayer Perception (MLP) networks and Alternating Decision Trees (ADtree) to classify objects as quasars or non-quasars using the cross-matched

* Supported by the National Natural Science Foundation of China.

results of a radio survey and a near infrared survey. Such a classification is helpful to preselect quasar candidates for large survey projects.

The structure of this paper is as follows: Section 2 presents the data collection and attribute selection. Section 3 gives a brief introduction of BBN, MLP and ADTree. The procedure to obtain the classifiers and the classification results are presented in Section 4. Section 5 summarizes and concludes the present work.

2 DATA SAMPLE AND CHOSEN ATTRIBUTES

We describe here the known near infrared, radio catalogs as follows. Table 1 summarizes the characteristics of the two surveys, FIRST and 2MASS.

Table 1 Summary of Catalog Characteristics

Survey	Wavelength	Sensitivity	Resolution (arcsec)	Number of Souces	Coverage Area
FIRST	21 cm	1 mJy	5	811 000	9033 deg^2
2MASS	j(1.25 μ m)	15.8 mag^a	0.5	470 992 970	41252.96 deg^2
	h(1.65 μ m)	15.1 mag^a			
	k(2.17 μ m)	14.3 mag^a			

^{a}For S/N$=10$

The Two Micron All Sky Survey (2MASS) project (Cutri et al. 2003) was designed to close the gap between our current technical capability and our knowledge of the near-infrared sky. 2MASS uses two new and highly-automated 1.3-m telescopes, one at Mt. Hopkins, Arizona, USA, and one at CTIO, Chile. Each telescope is equipped with a three-channel camera, each channel consisting of a 256×256 array of HgCdTe detectors, capable of observing the sky simultaneously at j (1.25 μm), h (1.65 μm), and k (2.17 μm), to a 3σ limiting sensitivity of 17.1, 16.4 and 15.3 mag in the three bands, respectively. The number of 2MASS point sources adds up to 470 992 970.

The survey FIRST (Faint Images of the Radio Sky at Twenty centimeters) began in 1993. It uses the VLA (Very Large Array, the National Radio Observatory (NRAO)) at a frequency of 1.4 GHz, and it is slated to 10 000 deg^2 of the North and South Galactic Caps, to a sensitivity of about 1mJy with an angular resolution of about 5 arcsec. The images produced by an automated mapping pipeline have pixels of 1.8 arcsec, a typical rms of 0.15 mJy, and a resolution of 5 arcsec; the images are available on the Internet (see the FIRST home page at *http://sundog.stsci.edu/* for details). The source catalogue is derived from the images. A new catalog (Becker et al. 2003) of the FIRST survey has been released that includes all taken from 1993 through September 2002, and contains about 811 000 sources covering 8422 deg ^{2} in the north galactic cap and 611 deg^2 in the south galactic cap. The new catalog and images are accessible via the FIRST search engine and the FIRST cutout server.

The 12th edition catalogue of quasars and active nuclei (Cat. VII/248, Véron-Cetty & Véron 2006) is an update of the previous versions, which now contains 85221 quasars, 1122 BL Lac objects and 21737 active galaxies (including 9628 Seyfert 1s), almost doubling the number listed in the 11th edition. As in the previous editions there is no information on the properties of X-ray absorption lines, but absolute magnitudes are given assuming $H_0 = 50$ km s^{-1} Mpc^{-1} and $q_0 = 0$. In this edition the 20 cm radio flux is listed when available, in place of the 11 cm flux.

The Tycho-2 Catalogue (Cat. I/259, Hog et al. 2000) is an astrometric reference catalogue containing positions and proper motions as well as two-color photometric data for the 2.5 million brightest stars in the sky. The Tycho-2 positions and magnitudes are based on precisely the same observations as the original Tycho Catalogue (hereafter Tycho-1; see Cat. I/239) collected by the star mapper of the ESA Hipparcos satellite, but Tycho-2 is much bigger and slightly more precise, owing to a more advanced reduction technique. Components of double stars with separations down to 0.8 arcsec are included. Proper motions precise to about 2.5 mas yr^{-1} are shown.

We first positionally cross-matched the 2MASS catalogue with the FIRST catalogue within 5 arcsec radius, then crossed out the one-to-many entries and obtained 153135 one-to-one entries. Secondly, the entries were cross-matched with qso.dat of the Véron-Cetty & Véron (2006)'s catalog and the Tycho-2

catalog within 5 arcsec radius, respectively. Similarly not considering the one-to-many entries, we obtained 2389 quasars and 1353 stars from the 2MASS and FIRST catalogues. The chosen attributes from different bands are $\log F_{\rm peak}$ ($F_{\rm peak}$: peak flux density at 1.4 GHz), $\log F_{\rm int}$ ($F_{\rm int}$: integrated flux density at 1.4 GHz), $f_{\rm maj}$ (fitted major axis before deconvolution), $f_{\rm min}$ (fitted minor axis before deconvolution), $f_{\rm pa}$ (fitted position angle before deconvolution), $j-h$ (near infrared index), $h-k$ (near infrared index), $k+2.5\log F_{\rm int}$, $k+2.5\log F_{\rm peak}$, $j+2.5\log F_{\rm peak}$ and $j+2.5\log F_{\rm int}$. To see the statistical properties of this sample, the mean values of the parameters are listed in Table 2. Meanwhile the distributions of the parameters are shown in Figure 1. As shown by Table 2, some mean values have rather large scatters. The values of $\log(F_{\rm peak})$, $\log(F_{\rm int})$, $k+2.5\log F_{\rm int}$, $k+2.5\log F_{\rm peak}$, $j+2.5\log F_{\rm peak}$ and $j+2.5\log F_{\rm int}$ for quasars are obviously bigger than those of stars. This means that quasars are generally stronger radio emitters than stars. In addition, the values of $j-h$ and $h-k$ of quasars are larger than those of stars, i.e. quasars are redder than stars. Moreover, Table 1 and Figure 1 indicate that $f_{\rm maj}$, $f_{\rm min}$ and $f_{\rm pa}$ are unimportant for discriminating quasars from stars while the other attributes are useful. Therefore, in the following we pick out quasars from stars considering two situations: situation 1 (S1), with all attributes and situation 2 (S2), without $f_{\rm maj}$, $f_{\rm min}$ and $f_{\rm pa}$.

Table 2 Mean Values of Parameters for the Samples

Parameters	Stars	Quasars
$\log(F_{\rm peak})$	0.46± 0.46	1.12± 0.87
$\log(F_{\rm int})$	0.55± 0.49	1.18± 0.91
$f_{\rm maj}$	7.22± 2.95	6.76± 2.93
$f_{\rm min}$	5.51± 1.28	5.51± 1.16
$f_{\rm pa}$	92.16± 59.29	87.61± 62.22
$j-h$	0.41± 0.52	0.64± 0.29
$h-k$	0.13± 0.37	0.61± 0.37
$k+2.5\log(F_{\rm peak})$	10.94± 2.50	17.69± 2.38
$k+2.5\log(F_{\rm int})$	11.16± 2.56	17.85± 2.46
$j+2.5\log(F_{\rm peak})$	11.43± 2.60	18.95± 2.35
$j+2.5\log(F_{\rm int})$	11.70± 2.65	19.10± 2.42

3 MODEL SELECTION

We used three methods to separate quasars from stars: BBN, MLP networks and ADTree. BBN has been used for the classification of variable stars (López et al. 2006). MLP and ADTree have been successfully used for the classification of multiwavelength data (see Zhang et al. 2004; Zhang et al. 2005; Zhang & Zhao 2004).

3.1 BBN

BBN is a powerful knowledge representation and reasoning tool under conditions of uncertainty which is defined by two components. The first is a direct acyclic graph, where each node represents a random variable and each arc represents a probabilistic dependence (Pearl 1988; Neapolitan 1990; Han & Kamber 2001). If an arc is drawn from a node Y to a node Z, then Y is a parent or immediate predecessor of Z, and Z is a descendent of Y. Each variable is conditionally independent of its nondescendents in the graph, given its parents. The variables may be discrete or continuous-valued. They may correspond to actual attributes given in the data or to hidden variables believed to form a relationship. The second component consists of one conditional probability table (CPT). The CPT for a variable Z specifies the conditional distribution $P(Z|{\rm Parent}(Z))$, where parent (Z) are the parents of Z. The joint probability of any n-tuple $(z_1, ..., z_n)$ corresponding to the variables or attributes $Z_1, ...Z_n$ is computed by

$$P(z_1, ..., z_n) = \prod_{i=1}^{n} P(z_i|{\rm Parent}(Z_i)),$$

where the values for $P(z_i|{\rm Parents}(Z_i))$ correspond to the entries in the CPT for Z_i. A node within the network can be selected as an output node, representing a class label attribute. There may be more than one

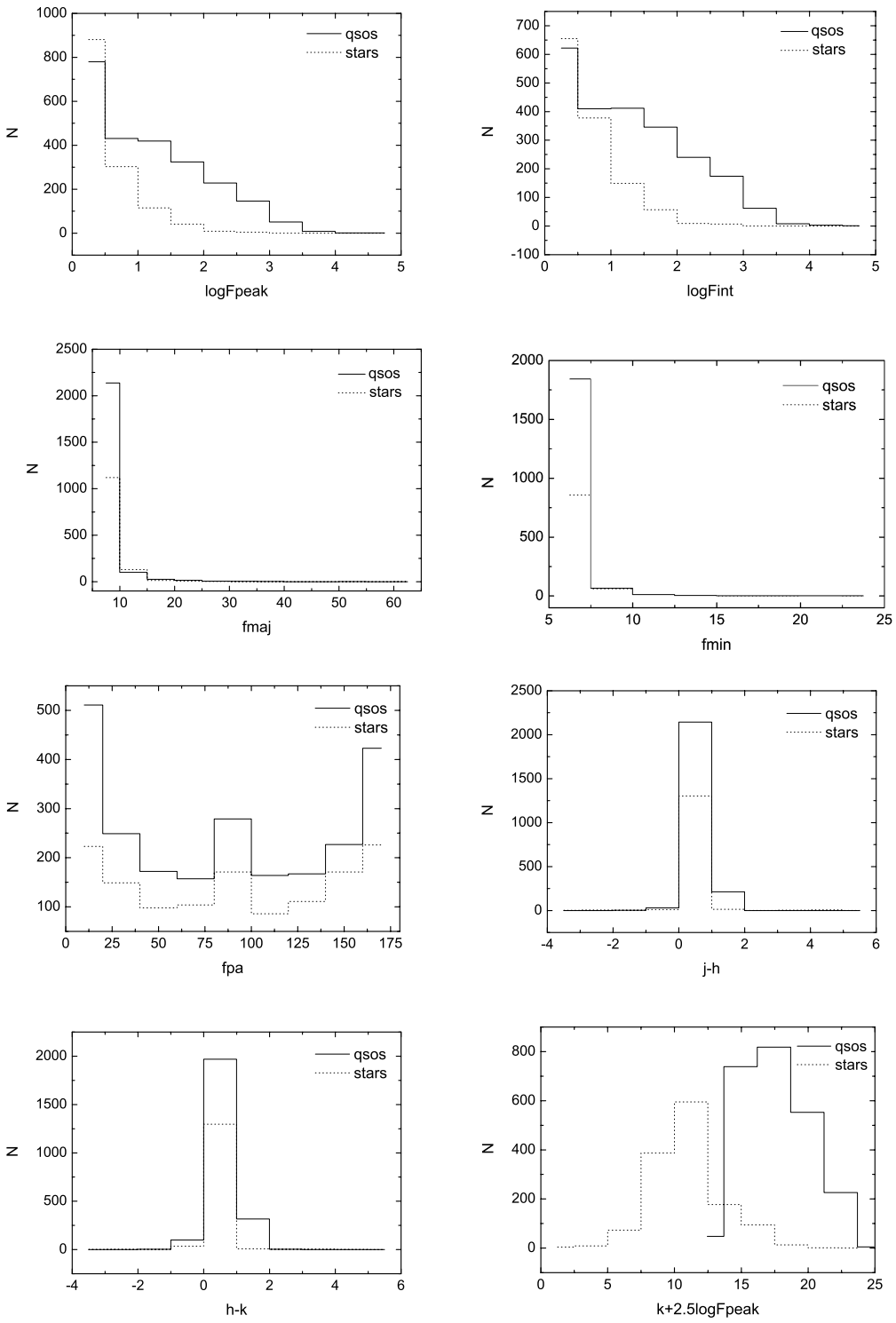

Fig. 1 Results of analysis of the sample for quasars (solid line) and stars (dotted line).

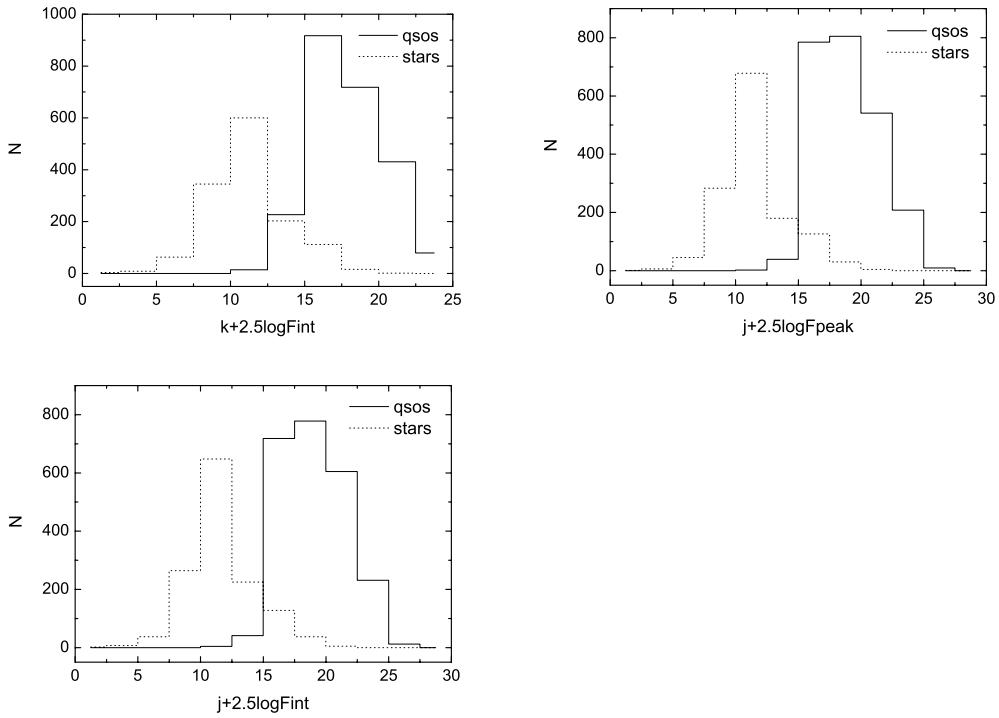

Fig. 1 –Continued.

output node. Inference algorithms for learning can be applied on the network. The classification process is such that, rather than returning a single class label, it can return a probability distribution for the class label attribute.

3.2 MLP Networks

MLP network is one of the most widely applied and investigated artificial neural network model. MLP networks have been applied successfully to solve some difficult and diverse problems. Training them in a supervised manner with a highly popular algorithm known as the error back-propagation (BP) has become very popular. The algorithm is based on an error-correction learning rule. MLP network model consists of a network of processing elements or nodes arranged in layers. Typically, it comprises three or more layers of processing nodes: an input layer which accepts the input variables used in the classification procedure, one or more hidden layers, and an output layer with one node for one class. In fact, a network with just two hidden units using the tanh function often fits the data quite well. The fit can be further improved by adding yet more units to the hidden layer. However, having too large a hidden layer - or too many hidden layers - can degrade the network's performance. In general, one should not use more hidden units than is necessary to solve a given problem.

One way to ensure this is to start training with a very small network. If gradient descent fails to find a satisfactory solution, then we expand the network by adding a hidden unit, and repeat the process. MLP network is a general-purpose, flexible, nonlinear model. It has been shown that, given enough hidden units and enough data, the MLP can approximate virtually any function to any desired accuracy. In other words, any function can be expressed as a linear combination of tanh functions: tanh is a universal basis function. Many functions form a universal basis; the two classes of activation functions commonly used in neural networks are the sigmoidal (S-shaped) basis functions (to which tanh belongs) and the radial basis functions. MLPs are valuable tools when one has little or no knowledge about the form of the relationship between input vectors and their corresponding outputs. Examples of applications of MLP networks in astronomy can be found in Vanzella et al.(2004). An introduction to Neural Networks is presented by Sarle (1994a), and

to MLP by Bailer-Jones et al.(2001) and Sarle (1994b). A comprehensive treatment of feed-forward neural networks is provided by Bishop (1995).

3.3 ADTree

ADTree is a generalization 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 which are included in a multi-path is the classification which the tree associates with the instance. The principle of the algorithm is presented in Freund & Mason (1999).

4 EXPERIMENTS AND RESULTS

Our experiments were done with the WEKA machine learning package (Witten & Frank 2005). In the process of experimenting, the default configurations of BBN, MLP and ADTree are used. The computer used in this effort was a PC with a 3.4 GHz Pentium 4 and CPU 1 GB memory. The operating system was Microsoft Windows XP. Here we use 10-fold cross-validation to evaluate the different accuracy of different models for this database. By comparing the accuracy of the classification and time taken to build the model, we try to compare the efficiency and effectiveness of the models.

4.1 Cross-Validation

Cross-validation is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis. K-fold cross-validation is one important cross-validation method. The data are divided into k subsets of (approximately) equal size. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. Then the average error across all k trials is computed. Cross-validation is often used for choosing among various models, such as different network architectures. For example, one might use cross-validation to choose the number of hidden units, or one could use cross-validation to choose a subset of the inputs (subset selection).

4.2 Results

Using a 10-fold cross-validation method we found the classification accuracy achieved with the different algorithms. The results are shown in Tables 3–6. Here MLP comprises a three-layer topology, i.e. it includes one input layer, one hidden layer and one output layer. Applying ADTree technique on the two samples, the total number of nodes is 31 and the number of predictor nodes is 21. For any algorithm, the accuracy of quasars and stars is more than 88.0%. For the sample S1, the number of correctly classified instances is 3524 for BBN, 3579 for MLP and 3553 for ADTree; as shown in Table 6, the corresponding overall accuracy for BBN, MLP and ADTree amounts to 94.17%, 95.64% and 94.95%, respectively; the running times to build the models is 0.34 s, 25.14 s and 1.25 s, respectively. Similarly, for the sample S2, the correctly classified instances for BBN, MLP and ADTree are 3531, 3585 and 3562; and Table 6 shows that the corresponding overall accuracies are respectively 94.36%, 95.80% and 95.19%; and the running times to build the models are respectively 0.28 s, 19.23 s and 0.86 s. From the results, we conclude that BBN, MLP and ADTree are all feasible for separating quasars from stars if only accuracy is considered. When only considering the running time, BBN is the fastest, and ADTree is faster than MLP. If considering both the accuracy and time, ADTree is the best. Tables 3–6 also indicate that both the accuracy and the speed of model building in all three algorithms are better in S2 than in S1. This fact clearly shows that the effectiveness and efficiency of these models with feature selection are a little better than those without feature selection. In addition, the classification results indicate that it is applicable to preselecting quasar candidates from the 2MASS and FIRST survey catalogues. The classifiers trained by these methods can be used to classify the unclassified sources.

Table 3 Classification Result for BBN with Different Samples

Sample	S1		S2	
classified↓ known→	stars	quasars	stars	quasars
stars	1190	55	1190	48
quasars	163	2334	163	2341
Accuracy	88.0%	97.7%	88.0%	98.0%

Table 4 Classification Result for MLP with Different Samples

Sample	S1		S2	
classified↓known→	stars	quasars	stars	quasars
stars	1220	30	1220	24
quasars	133	2359	133	2365
Accuracy	90.0%	98.7%	90.2%	99.0%

Table 5 Classification Result for ADTree with Different Samples

Sample	S1		S2	
classified↓known→	stars	quasars	stars	quasars
stars	1194	30	1200	27
quasars	159	2359	153	2362
Accuracy	88.2%	98.7%	88.7%	98.9%

Table 6 Accuracy and Time to Built Model of Different Methods for Different Samples

Sample	S1		S2	
Method	Accuracy	Time	Accuracy	Time
BBN	94.17%	0.34 s	94.36%	0.28 s
MLP	95.64%	25.14 s	95.80%	19.23 s
ADTree	94.95%	1.25 s	95.19%	0.86 s

5 CONCLUSIONS

Survey data are important source of information. By classification techniques, we can extract a large volume of information from the raw data. Here we analyzed a sample and compared the results with and without feature selection. We found that when algorithms with feature selection are applied, the accuracy and the speed to build models are both better than the results without feature selection. Clearly appropriate feature selection may improve the effectiveness and efficiency of classifiers. For the given task, BBN, MLP and ADTree models when applied to the present sample, achieved higher accuracy of more than 94.0%. If we only take the accuracy into account, then BBN, MLP and ADTree gave comparable performance. However, BBN classifier has the fastest speed, especially when applied to large databases. When both accuracy and speed are considered, ADTree comes out best. In conclusion, these three algorithms are robust and efficient methods for solving the classification problems we are facing in astronomy.

The classifiers obtained by these methods may be used to preselect source candidates of interest to the astronomers. These techniques may be used on other types of data, such as spectral data and image data. Moreover, they are also applicable to other aspects, for example star parameter measurement, redshift estimation of galaxies and quasars, morphology classification of galaxies. As the quantity, quality and complexity of astronomical data continue to improve, and the number of features continually rising, selecting appropriate models and training the classifiers efficiently, as well as feature selection methods in general, are challenging work for future research. Especially faced with large and multi-wavelength sky surveys, automated methods not only reduce the astronomer's chore, but also improve the efficiency of astronomers and high-cost telescopes; additionally, effective feature selection methods reduce the dimensionality of the parameter space, and improve the efficiency and effectiveness of automated classification algorithms, thereby making it possible for the application of some methods only used in low dimensional spaces. Successful application of data mining in astronomical databases is the catalyzer of finding unusual, rare or unknown

objects and phenomenon. Especially, clustering analysis and outlier finding algorithms can facilitate class discovery in astronomy.

Acknowledgements We are very grateful to an anonymous referee for his/her helpful comments and suggestions. Many thanks go to Gao Dan, for her work of data processing. This research has 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 work is funded by the National Natural Science Foundation of China (NSFC) under Grants 10473013 and 90412016.

References

Bailer-Jones C. A. L., Gupta R., Singh H. P., 2001, An introduction to artificial neural networks Proc. of the Workshop on Automated Data Analysis in Astronomy, IUCAA, Pune, India, October 9.12, 2000 (arXiv:astro-ph/0102224)

Becker R. H., Helfand D. J., White R. L. et al., 2003, VizieR On-line Data Catalog: VIII/71

Bishop C. M., 1995, Neural Networks for Pattern Recognition, Oxford: Oxford University Press

Cutri R. M., Skrutskie M. F., van Dyk S. et al., 2003, VizieR On-line Data Catalog: II/246

Fayyad U. M., Piatetsky-Shapiro G., Smyth P. et al., 1996, Advances in Knowledge Discovery and Data mining. California: AAAI/MIT Press

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

Han J. W., Kamber M., 2001, Data Mining: Concepts and Techniques, Higher Education Press

Hog E., Fabricius C., Makarov V. V. et al., 2000, A&A, 355, L27

López M., Bielza C., Sarro L. M., 2006, Astronomical Data Analysis Software and Systems XV ASP Conference Series, eds. by Carlos Gabriel, Christophe Arviset, Daniel Ponz, and Enrique Solano. San Francisco: Astronomical Society of the Pacific, 351, 161

Neapolitan R. E., 1990, Probabilistic reasoning in expert systems: theory and algorithms, John Wiley & Sons

Pearl J., 1988, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann

Sarle W. S., 1994, Neural Networks and Statistical Models, Proc. of the Nineteenth Annual SAS Users Group International Conf., Cary, NC: SAS Institute, 1538 (ftp://ftp.sas.com/pub/neural/neural1.ps)

Sarle W. S., 1994, in: Neural Network Implementation in SAS Software, SAS Institute Inc., Proc. of the Nineteenth Annual SAS Users Group International Conf., Cary, NC: SAS Institute Inc., p.1551 (ftp:// ftp.sas.com/pub/neural/neural2.ps)

Vanzella E., Cristiani S., Fontana A. et al., 2004, A&A, 423, 761

Véron-Cetty M. P., Véron P., 2006, A&A, 455, 733

Witten I. H., Frank E.,: Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005

Zhang Y., Luo A., Zhao Y., 2004, Ground-based Telescopes. eds. by Oschmann, Jacobus M., Jr. Proceedings of the SPIE, 5493, 483

Zhang Y., Zhao Y., 2004, A&A, 422, 1113

Zhang Y., Zheng H., Zhao Y., 2005, IAU Colloquium Proceedings of the International Astronomical Union 199, eds. P. R. Williams, C.-G. Shu and B. Menard., Cambridge: Cambridge University Press, p.481