An automated classification algorithm for multi-wavelength data

Yanxia Zhang [*a], Ali Luo [*a], and Yongheng Zhao [*a]

[*a]National Astronomical Observatories, Chinese Academy of Sciences, China.

ABSTRACT

The important step of data preprocessing of data mining is feature selection. Feature selection is used to improve the performance of data mining algorithms by removing the irrelevant and redundant features. By positional cross-identification, the multi-wavelength data of 1656 active galactic nuclei (AGNs), 3718 stars, and 173 galaxies are obtained from optical (USNO-A2.0), X-ray (ROSAT), and infrared (Two Micron All- Sky Survey) bands. In this paper we applied a kind of filter approach named ReliefF to select features from the multi-wavelength data. Then we put forward the naive Bayes classifier to classify the objects with the feature subsets and compare the results with and without feature selection, and those with and without adding weights to features. The result shows that the naive Bayes classifier based on ReliefF algorithms is robust and efficient to preselect AGN candidates.

Keywords: Feature selection, Classification, Astronomical databases: miscellaneous, Catalogs, Methods: Data Analysis, Methods: Statistical

1. INTRODUCTION

Astronomy is entering a new data avalanche era as multiple, large area, digital sky surveys in production. Many of the catalogs will cover different wavebands, from the radio to optical/infared all the way to the X-rays, even γ rays. The amount of catalogued data is measured by Terabytes, even Petabytes. The aggregation and federation of the multi-wavelength datasets, however, is a challenging task to astronomers. To solve the problems, the international virtual observatory alliance establishes. Data mining is a major factor of virtual observatory. Since celestial objects generally radiate energy over an extremely wide range of wavelengths. Each of these observations carries important information about the nature of the objects. Consequently, in order to study the properties of objects in detail, we need to consider the various parameters from different bands and classify objects in the multidimensional parameter space. However,the answers to the questions are based not only on the improvement of the technologies of telescopes and detectors, but also on the development of well-automated classification tools.

When data mining aims to solve larger, more complex tasks, it has become increasing important how to deal with the most relevant information in a potentially overwhelming quantity of data. With a data flood coming, the typical data set consists of $\sim 10^8 - 10^9$ sources with $\sim 10^2$ measured attributes each, i.e., a set of $\sim 10^9$ data vectors in a 100-dimensional parameter space. In many cases, the majority of the features handled by data mining scheme are irrelevant or redundant. From a data perspective, it is necessary to perform pre-processing and then select a subset of features. A successful choice of features provided to a classifier can improve its accuracy, save the computation time, and simplify its result.

John et al. (1994) proposed two models to select a "good" set of features under some objective function. The feature filter model assumes filtering the features before applying an induction algorithm, while the wrapper model uses the induction algorithm itself to evaluate the features. In detail, the wrapper employs a statistical re-sampling technique with the actual target learning algorithm to estimate the accuracy of feature subsets. This method has proved useful but is very slow to execute because the learning algorithm is called repeatedly.

*contact: zyx@lamost.bao.ac.cn; lal@lamost.bao.ac.cn; yzhao@lamost.bao.ac.cn; phone 8610-64841693; fax 8610-64878240; Datun Road 20A, Chao Yang District, National Astronomical Observatories,Chinese Academy of Sciences, China 100012

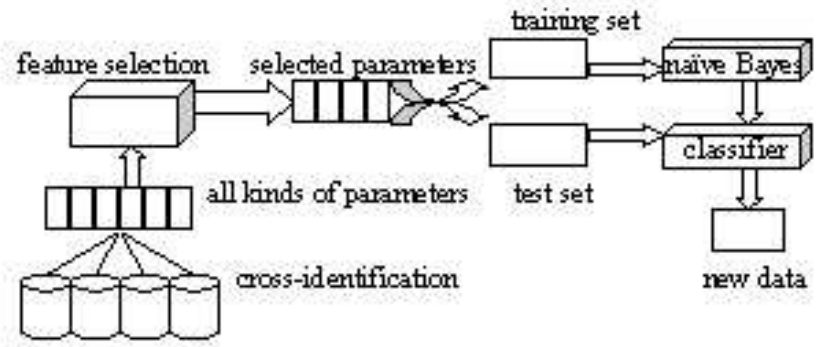

Figure 1. the scheme of classification

As a result, wrappers do not scale well to large datasets containing many features. In contrast, the filter operates independently of any learning algorithm. The undesirable features are filtered out of the data before induction commences. Filters typically make use of all the available training data when selecting a subset of features. Some look for consistency in the data, namely they note when every combination of values for a feature subset is associated with a single class label (Almuallim and Dietterich, 1992). Another method (Koller and Sahami, 1996) eliminates features whose information content is subsumed by some number of the remaining features. Still other methods attempt to rank features according to a relevancy score (Kira and Rendell, 1992; Holmes and Nevill-Manning, 1995). Filters have proven to be much faster than wrappers and hence can be applied to large data sets containing many features. Their general nature allow them to be used with any learner, unlike the wrapper, which must be re-run when switching from one learning algorithm to another. As for the review of feature selection can refer to Liu and Motoda (1998).

Quasar candidate selection methods employed by previous surveys include radio selection, color selection, slitless spectroscopy (SS) selection, X-ray 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 color and SS selection. Zhang et al. (2002, 2003a, 2003b) explored automated classification methods, Learning Vector Quantization (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.

In this paper, we introduce an efficient feature selection algorithm, called ReliefF, which evaluates each attribute by its ability to distinguish among instances that are near each other. Their selection criterion, the feature relevance, is applicable to numeric and nominal attributes. The threshold of relevancy is determined statistically by using Chebyshev's inequality, which is not sharp enough making a clear distinction between relevant and non-relevant features. With the subset of features obtained by ReliefF, we divide the data into two parts: one as the training set, another as the test set. Then we use the training set to train the naive Bayes and get the naive Bayes classifier. With the test set to test the classifier, we apply the classifier to classify the new data if the classifier is good. The whole scheme is described in Figure 1. We compare the results with and without feature selection, and those with and without adding weight to features.

2. METHODOLOGY

2.1. ReliefF

The ReliefF algorithm (Kononenko, 1994; Robnik-Šikonja & Kononenko, 2003) (See Figure 2) is the extension of the Relief algorithms. The original Relief can deal with nominal and numerical attributes, however it cannot deal with incomplete data and is limited to two-class problems. In contrast, the ReliefF algorithm solves these and

other problems, moreover it is not limited to two class problems, is more robust and can deal with incomplete and noisy data. ReliefF randomly selects an instance R_i (line 3), then search for k of its nearest neighbors from the same class, called nearest hits H_j (line 4), and also k nearest neighbors from each of the different classes, called nearest misses $M_j(C)$ (lines 5 and 6). It updates the quality estimation $W[A]$ for all attributes A depending on their values for R_i, hits H_j and misses $M_j(C)$ (lines 7, 8 and 9). The update formula is similar to that of Relief, except that we average the contribution of all the hits and all the misses. The contribution for each class of the misses is weighted with the prior probability of that class $P(C)$ (estimated from the training set). Since we want the contributions of hits and misses in each step to be in [0,1] and also symmetric, we have to ensure that misses' probability weights sum to 1. As the class of hits is missing in the sum, we have to divide each probability weight with factor $1 - P(class(R_i))$ (which represents the sum of probabilities for the misses' classes). The process is repeated for m times, where m is a user-defined parameter.

Function $diff(A, I_1, I_2)$ calculates the difference between the values of the attribute A for two instances I_1 and I_2. For nominal attributes, it was originally defined as:

$$diff(A, I_1, I_2) = \begin{cases} 0; value(A, I_1) = value(A, I_2) \\ 1; otherwise \end{cases}$$

For numerical attributes as:

$$diff(A, I_1, I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{max(A) - min(A)}$$

To deal with incomplete data, the diff function needs to be changed. Missing values of attributes are treated probabilistically. We calculate the probability that two given instances have different values for given attribute conditioned over class value:
– if one instance (eg., I_1) has unknown value:

$$diff(A, I_1, I_2) = 1 - P(value(A, I_2)|class(I_1)) \tag{1}$$

– if both instances have unknown value:

$$diff(A, I_1, I_2) = 1 - \sum_{V}^{\#values(A)} (P(V|class(I_1)) \times P(V|class(I_2))) \tag{2}$$

Conditional probabilities are approximated with relative frequencies from the training set.

Algorithm ReliefF

Input: for each training instance a vector of attribute values and the class value
Output: the vector W of estimations of the qualities of attributes
1. set all weights $W[A] := 0.0$;
2. for $i := 1$ to m do begin
3. randomly select an instance R_i;
4. find k nearest hits H_j;
5. for each class $C \neq class(R_i)$ do
6. from class C find k nearest misses $M_j(C)$;
7. for $A := 1$ to a do
8. $W[A] := W[A] - \sum_{j=1}^{k} diff(A, R_i, H_j)/(m \cdot k) +$
9. $\sum_{C \neq class(R_i)} \sum_{j=1}^{k} diff(A, R_i, M_j(C)]/(m \cdot k)$;
10. end;

Figure 2. Pseudo code of Relief algorithm

2.2. Naive Bayes classifiers

The naive Bayes classifiers assign the most likely class to a given example described by its feature vector. The classifiers assume that the effect of an variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. It is made to simplify the computation and in this sense considered to be "naive. This assumption is a fairly strong assumption and is often not applicable. However, bias in estimating probabilities often may not make a difference in practice – it is the order of the probabilities, not their exact values, that determine the classifications. In practice the naive Bayes classifiers are often comparable in performance with more sophisticated classifiers such as decision tree and neural network classifiers (Hilden, 1984; Langley, Iba, & thompson, 1992; Friedman, Geiger, & Goldszmidt, 1997; Domingos & Pazzani, 1997). Naive Bayes has proven effective in many practical applications, including text classification, medical diagnosis, and systems' performance management (Domingos & Pazzani, 1997; Mitchell, 1997; Hellerstein, thathachar, & Rish, 2000). They have also exhibited high accuracy and speed when applied to large databases.

Here a more technical description of the naive Bayes is given. Let X be the data record (case) whose class label is unknown. Let H be some hypothesis, such as "data record X belongs to a specified class C." For classification, we want to determine $P(H|X)$ – the probability that the hypothesis H holds, given the observed data record X.

$P(H|X)$ is the posterior probability of H conditioned on X. In contrast, $P(H)$ is the prior probability, or a priori probability of H. The posterior probability, $P(H|X)$, is based on more information (such as background knowledge) than the prior probability, $P(H)$, which is independent of X.

Similarly, $P(X|H)$ is posterior probability of X conditioned on H. $P(X)$ is the prior probability of X. Bayes theorem is useful in that it provides a way of calculating the posterior probability, $P(H|X)$, from $P(H)$, $P(X)$, and $P(X|H)$. Bayes theorem is

$$P(H|X) = P(X|H) \times P(H)/P(X) \tag{3}$$

3. DATA

The ROSAT Bright Source (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 (FSC) includes 10,5924 sources. A catalogue of quasars and active nuclei (Véron-Cetty & Véron, 2000) contains 13214 quasars, 462 BL Lac objects and 4428 active galaxies (of which 1711 are Seyfert 1).

We positionally cross-identify the Véron 2000 catalog with the ROSAT Bright Source Catalog (RASS/BSC) and Faint Source Catalog (RASS/FSC) X-ray sources, and then cross-identify the result with optical sources in the USNO A-2.0 catalog. 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 Third Reference Catalogue of Bright Galaxies (RC3; de Vaucouleursget et al. 1991) to obtain 3718 stars and 173 normal galaxies from optical, X-ray and infrared bands. The chosen attributes from different bands are $B-R$ (optical index), $B+2.5log(CR)$, $lgCR$ (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), $J+2.5log(CR)$.

Wei et al. (1999) chose $logC \geq 0.4R+4.9$ as a criterion to preselect active galactic nucleus (AGN) samples, where C is the X-ray count rate and R is the R magnitude. According to the results of the Einstein Medium Sensitivity Survey (EMSS; Stocke et al. 1991), the X-ray to optical flux ratio F_X/F_{opt} was found to be very different for different classes of X-ray emitters. Motch et al. (1998) stated that for source classification, the most interesting parameters are flux ratios in various energy bands, including the conventional X-ray hardness ratios, F_X/F_{opt} ratios, and optical colors. They also found that, although stars and AGNs have similar X-ray colors, their mean X-ray-to-optical ratios are obviously quite different, and they are well separated in the HR1/2 versus F_X/F_{opt} diagram. Cataclysmic variables exhibit a large range of X-ray colors and F_X/F_{opt} ratios and can be somewhat confused with both AGNs and the most active part of the stellar population. However, the addition of a B-V or U-B optical index would allow further distinguishing between these overlapping populations.

Pietsch et al. (1998) also used a conservative extent criterion as an indicator that the X-ray emission does not originate from a nuclear source. As a result, based on the optical classification, X-ray characteristics such as hardness ratios and the extent parameter, the infrared classification and so on (Stocke et al. 1991; Motch et al. 1998; Pietsch et al. 1998), the present investigation is plausible. For different classes of objects, the different distributions of attributes are helpful to classify the objects.

4. RESULTS AND DISCUSSION

4.1. Results

We present the result of attribute estimation on the multi-wavelength data based on common description by the 10 attributes. ReliefF separates the important attributes from unimportant ones. The attributes with more values convey more information. The rank of importance of these attributes in sequence is $B + 2.5log(CR)$, $J + 2.5log(CR)$, $B - R$, $HR2$, $H - K$, ext, $J - H$, $lgCR$, $HR1$, $extl$. The estimation of these attributes is given as follows:

B+2.5log(CR)	J+2.5log(CR)	B-R	HR2	H-K
0.04207	0.03838	0.03230	0.01332	0.01011

ext	J-H	lgCR	HR1	extl
0.00716	0.00317	0.00213	0.00185	0.00096

From the above result, it shows that $B + 2.5log(CR)$, $J + 2.5log(CR)$, $B - R$, $HR2$, $H - K$ and ext are the good attributes to carrying the most information to discriminate AGNs from stars and normal galaxies. The rest attributes are less important. Though feature selection by ReliefF, we choose the good attributes as the feature subset for classification. To check the performance of classification with the feature subset, we compare two situations: with the feature subset and with the full set of features, as the input of the naive Bayes classifier, respectively. Randomly dividing the sample into two sets: one for training set and another for test set, we use the training set to train and get the naive Bayes classifier. After that, we employ the test set to evaluate the performance of the classifier. The classification results are shown in Tables 1-2, separatively. The total accuracy of the two situations is 97.9% and 97.0%, respectively.

Table 1. The classification result with the feature subset

classified↓known→	AGNs	non-AGNs
AGNs	833	44
non-AGNs	14	1883
accuracy	98.3%	97.7%

In addition, the ReliefF algorithm assigns a weight (importance) to each feature. So we can use their weights directly. Adding different weights to corresponding features, we randomly divide the sample to two parts: one for training set and another for test set. Just like above steps, we get the naive classifier and give the classification result in Table 3. The total accuracy add up to 97.6%.

4.2. Discussion

Facing various large sky surveys, we need improving efficiency of high-costly telescopes and developing automated and robust approaches to preselect AGN candidates or other source candidates. The naive Bayes

Table 2. The classification result with the full set of features

classified↓known→	AGNs	non-AGNs
AGNs	798	74
non-AGNs	10	1892
accuracy	98.8%	96.2%

Table 3. The classification result with the weighted attributes

classified↓known→	AGNs	non-AGNs
AGNs	794	21
non-AGNs	45	1914
accuracy	97.4%	97.7%

classifier based on RliefF algorithm gives high accuracy (higher than 97%) of classifying AGNs from stars and normal galaxies with multi-wavelength data. So this method can be used for large sky survey, such as Chinese LAMOST.

From the above classification result, we see that RliefF algorithm separates the important attributes from unimportant ones. Tables 1-2 present that the classification result with feature subset selection by RliefF is better than that without feature selection. Moreover the classification result with the weighted features is also better than that with the unprocessed full set of features, as shown by Table 2 and Table 3. Obviously the naive Bayes classifier based on RliefF shows better performance than that independent on RliefF. Consequently ReliefF is an efficient and robust feature selection algorithm, meanwhile, is a good feature weighting approach.

For ReliefF algorithm, the goodness of a feature subset can be assessed only depending on the intrinsic properties of the data. It ignores the induction algorithm to assess the merits of a feature subset and performs the feature selection before applying the learning algorithm. Looking just at the data and considering the target concept to be learned. The learning algorithm constructs the concept using the set of selected features, ignoring the others. By removing or decreasing irrelevant information and redundant information, ReliefF algorithm improves the performance of the naive Bayes classifier.

Feature extraction and feature selection are important steps before data mining. Feature extraction methods include projection pursuit (see Friedman, 1987), factor analysis (see Kim and Mueller, 1978) and principal component analysis (see Dunteman, 1989; Zhang, 2003a), etc. Feature selection methods include wrapper approaches, filter approaches and embedded approaches, etc. The techniques are complementary in their goals: feature selection leads to savings in measurement cost and the selected features retain their original physical interpretation. On the other hand, the transformed features obtained by feature extraction techniques may provide a better discriminatory ability than the best selected subset, but these features fail in retaining the original physical interpretation and may not have a clear meaning. According to different tasks and demands, we choose appropriate feature extraction, feature selection and feature weighting approaches.

Despite its unrealistic independence assumption, the naive Bayes classifier is surprisingly effective in classify the multi-wavelength data since its classification decision may often be correct even if its probability estimates are inaccurate. When data are preprocessed by ReliefF, the performance of naive Bayes classifier increases. Obviously, a deeper understanding of data characteristics that affect the performance of naive Bayes is still required.

5. CONCLUSION

In this paper, we proposed a novel automated classification method, the naive Bayes classifier based on ReliefF, introduced an efficient way (ReliefF) of analyzing feature redundancy and assigning weight to features according

to the amount of information the features convey. The feature selection results are verified by applying the naive Bayes classifier to data with and without feature selection, and with and without weighting features. Our approach shows its efficiency and effectiveness in dealing with high dimensionality data for classification. With the quantity, quality and complexity of data improving, more effective and efficient classification techniques are required. The successful techniques may be used in all kinds of classification tasks, such as preselecting source candidates and object classification, and also be used for other types of data, for example, photometric data and spectral data. Our further work will extend the method to work on higher dimensionality, develop more effective feature selection approaches, or combined the feature selection techniques with other classifiers. With more classification schemes in practice, the data mining toolkits of virtual observatory will be enriched.

ACKNOWLEDGMENTS

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.10273011.

REFERENCES

1. C. Motch, P. Guillout, F. Haberl, et al. "Identification of selected sources from the ROSAT Galactic Plane Survey. I.", A&AS, **132**, pp. 341-359, 1998.
2. D. Koller & M. Sahami. "Toward optimal feature selection". In Proc. 13th Intl. Conf. on Machine Learning, pp. 284-292, Bari, Italy, 1996.
3. G. H. Dunteman, Principal Components Analysis, Sage Publications, Inc., NewburyPark, CA, 1989.
4. G. Holmes & C. G. Nevill-Manning, "Feature Selection via the dicovery of Simple Classification Rules". In Proceedings of the International Symposium on Intelligent data Analysis, 1995
5. G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem", Proc. Machine Learning: Proc. of the 11th Int'l Conf., san Mateo, CA, pp. 121-129, 1994.
6. H. Almuallim & T. G. Dietterich, "Efficient Algorithms for Identifying Relevant Features", Proc. 9th Canadian Conf. on AI, pp. 38-45, 1992
7. H. Liu & H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer, Boston, 1998.
8. I., Kononenko, "Estimating attributes: analysis and extensions of Relief", In: L. De Raedt and F. Bergadano (eds): Machine Learning: ECML-94, pp.171-182, Springer Verlag, 1994.
9. J. H. Friedman, "Exploration Projection Pursuit", Journal of American Statistical Association, **82**, pp. 249, 1987.
10. J. Hellerstein, Jayram Thathachar, and I. Rish, Recognizing end-user transactions in performance management, In Proceedings of AAAI-2000, pp. 596-602, Austin, Texas, 2000.
11. J. Hilden, "Statistical diagnosis based on conditional independence does not require it", Comput. Biol. Med., **14(4)**, pp. 429-435, 1984.
12. J. T. Stocke, Simon L. Morris, I. M. Gioia, et al. "The Einstein Observatory Extended Medium-Sensitivity Survey. II - The optical identifications", ApJS, **76**, pp. 813-874, 1991.
13. J-O. Kim & C. W. Mueller, Factor Analysis: Statistical Methods and Practical Issues, Sage Publications, Inc., Beverly Hills, 1978.
14. J. Y. Wei, D. W. Xu, X. Y.Dong, J. Y. Hu, "An AGN sample with high X-ray-to-optical flux ratio from RASS. I. The optical identification", A&AS, **139**, pp. 575-599, 1999.
15. K. Kira & L. Rendell, A Practical Approach to Feature Selection. In Machine Learning: Proceedings of the Ninth International Conference. Morgan Kaufmann, 1992
16. M. Robnik-*Š*ikonja, I., Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF", Machine Learning Journal, **53**, pp. 23-69, 2003.
17. M. P. Véron-Cetty, P. Véron, ESO Scientific Report 19, 2000.
18. N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers", Machine Learning, **29**, pp. 131-163, 1997.

19. P. Domingos & M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss", Machine Learning, **29**, pp. 103-130, 1997.
20. P. C. Hewett, C. B. Foltz, and F. H. Chaffee, "The large bright quasar survey. 6: Quasar catalog and survey parameters", Astron.J. **109**, pp. 1498-1521, 1995.
21. P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifier", In Proceedings of the Tenth National Conference on Artificial Intelligence, pp. 399-406, San Jose, CA, 1992. AAAI Press.
22. Tom M. Mitchell, Machine Learning, McGraw-Hill, 1997.
23. W. Voges, B. Aschenbach, Th. Boller, et al., "The ROSAT all-sky survey bright source catalogue", A&A **349**, pp. 389-405, 1999.
24. W. Pietsch, K. Bischoff, Th. Boller, et al. "New active galactic nuclei detected in ROSAT All Sky Survey galaxies. I. Verification of selection strategy", A&A, **333**, pp.48-62, 1998.
25. Y.-X, Zhang, C.-Z., Cui, and Y.-H., Zhao, "Classification of AGNs from stars and normal galaxies by support vector machines", Astronomical Data Analysis II. Edited by Starck, Jean-Luc; Murtagh, Fionn D., Proc. SPIE, **4847**, pp. 371-378, 2002.
26. Y.-X., Zhang & Y.-H., Zhao, "Classification in Multidimensional Parameter Space: Methods and Examp les", PASP, **115**, pp. 1006-1018, 2003a.
27. Y.-X., Zhang & Y.-H., Zhao, "Learning Vector Quantization for Classifying Astronomical Objects", CHJAA, **3**, pp. 183-190, 2003b.