An automated classification algorithm for multi-wavelength data

Yanxia Zhang, Ali Luo, Yongheng ZHAO

National Astronomical Observatories, Chinese Academy of Sciences

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 173galaxies 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 naïve 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 robust and efficient to preselect AGN candidates.

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

1. MOTIVATION

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 classier. With the test set to test the classier, we apply the classier to classify the new data if the classier 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 weights to features.

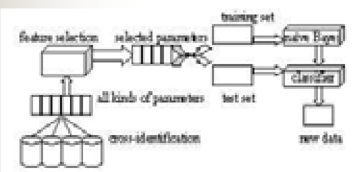

Figure 1. the scheme of classication

2. METHODOLOGY

2.1 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. $W[A] := W[A] - \Sigma_j diff(A,R_j,H_j)/(m \cdot k) +$
9. $\Sigma_{C \neq class(Ri)}\ \Sigma_j diff(A,R_j,M_j(C))/(m \cdot k)$, (j=1,2,....,k);
10. End;

2.2 Naive Bayes classifiers

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

$$P(H|X) = P(X|H) \times P(H)/P(X)$$

3. DATA

We positionally cross-identify the Veron 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 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, 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).

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 10attributes. 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	0.03898	0.03290	0.01332	0.01113

ext	J-H	lgCR	HR1	extl
0.00716	0.00317	0.00213	0.00195	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%

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%

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 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. Tables1-2 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 , factor analysis and principal component analysis, 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.

Email: zyx@bao.ac.cn; lal@bao.ac.cn; yzhao@lamost.org

National Astronomical Observatories, Chinese Academy of Sciences

China-VO

NAOC