Chin. J. Astron. Astrophys. Vol. 3 (2003), No. 2, 183–190
(*http://www.chjaa.org* or *http://chjaa.bao.ac.cn*)

Chinese Journal of
Astronomy and
Astrophysics

Learning Vector Quantization for Classifying Astronomical Objects *

Yan-Xia Zhang and Yong-Heng Zhao

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

Received 2002 June 28; accepted 2002 October 8

Abstract The sizes of astronomical surveys in different wavebands are increasing rapidly. Therefore, automatic classification of objects is becoming ever more important. We explore the performance of learning vector quantization (LVQ) in classifying multi-wavelength data. Our analysis concentrates on separating active sources from non-active ones. Different classes of X-ray emitters populate distinct regions of a multidimensional parameter space. In order to explore the distribution of various objects in a multidimensional parameter space, we positionally cross-correlate the data of quasars, BL Lacs, active galaxies, stars and normal galaxies in the optical, X-ray and infrared bands. We then apply LVQ to classify them with the obtained data. Our results show that LVQ is an effective method for separating AGNs from stars and normal galaxies with multi-wavelength data.

Key words: method: data analysis — method: statistical — catalogs

1 INTRODUCTION

With the vast amounts of data resulting from large digital sky surveys and archives, data measured in terabytes, and soon in petabytes, now becoming available, efforts are being made to develop algorithms for automatic classification. The advantages of automated procedures as compared to manual classification are obvious. Only a few experts are able to perform accurate manual classification, and it was therefore sought to "freeze" this expert knowledge into computer programs. Such programs would allow us to obtain objective classification by quantitative criteria, and handle much larger data sets. The latter issue has become ever more demanding, with ongoing and upcoming survey missions like SDSS, 2MASS, DPOSS, and LAMOST, etc. Hence, there is need for tools that provide efficient and robust methods of automatic classification of all detected objects.

A large amount of work has been dedicated to automatic object classification. Neural networks (NNs), over the years, have proven to be a very powerful tool capable of extracting reliable information and patterns from large amounts of data even in the absence of a model

* Supported by the National Natural Science Foundation of China.

describing the data (cf. Bishop 1995) in a wide range of applications: catalogue extraction (Andreon et al. 2000), star/galaxy classification (Odewahn et al. 1992; Naim et al. 1995; Mähönen & Hakala 1995; Bertin & Arnout 1996; Bazell & Peng 1998), galaxy morphology (Storrie-Lombardi et al. 1992; Lahav et al. 1996), classification of stellar spectra (Bailer-Jones et al. 1998; Allende et al. 2000; Weaver 2000). Recently an important and promising contribution was introduced by Andreon et al. (2000), covering a large number of neural algorithms.

Mähönen & Hakala (1995) also took a new approach to classification. They used a Kohonen self-organizing map (SOM), which is an unsupervised learning method, to distinguish images of stellar objects, nonstellar objects, and background. Two advantages of their approach are that the method is unsupervised and that it does not require data preprocessing. Now, supervised learning methods typically give better results than unsupervised methods. In this paper, we introduce a supervised neural network called learning vector quantization (LVQ) to classify multi-wavelength data, which is a supervised variation of the Kohonen self-organizing map (SOM), and of which Bazell & Peng (1998) pioneered applications in astronomy. LVQ shares the same network architecture as SOM, although it uses a supervised learning algorithm.

2 LEARNING VECTOR QUANTIZATION

Classification and pattern recognition are important and challenging issues in efficient analysis of large astronomical databases and will become even more important with the development of the International Virtual Observatory (IVO). Neural networks (NNs) classifiers have proved to be strong competitors in this field and especially in the discrimination of astronomical objects. The adopted learning vector quantization (LVQ) algorithm here is based on the LVQ_PAK routines developed at the Laboratory of Computer and Information Sciences, Helsinki University of Technology, Finland. The software can be obtained from www.cis.hut.fi/research/lvq_pak/.

According to learning process, neural networks are divided into two kinds: supervised and unsupervised. The difference between them lies in how the networks are trained to recognize and categorize the objects. In the unsupervised method, samples are input into the network and the network must determine the correlations between the objects and produce an output in the correct class for each input object. In essence, the unsupervised algorithm must have some internal means of differentiating objects in order to classify them. On the other hand, with the supervised learning method, the network is given input samples from a training data set, along with the current classification of each sample, and produces an output signifying its best guess for the classification of each input object. The network compares its output with the correct, or target output which was specified by the user along with the input data. The network then adjusts its internal components (connection weights) to make its output agree more closely with the target output. In this way the network learns the correct classification of its training data set. The network can then be presented with a test data set consisting of objects which it has never known, and its performance can be evaluated.

Learning vector quantization was also developed by Kohonen (1989) and is based on the self-organizing map (SOM) or Kohonen feature map (Kohonen 1989, 1990). SOM performs a mapping from an n-dimensional input vector onto a two-dimensional array of nodes usually displayed in a rectangular or hexagonal lattice. The mapping is performed in such a way as to preserve the topology of the input data. This means that input vectors, that are similar to each other in some sense, are mapped to neighboring regions of the two-dimensional output lattice.

Each node in the output lattice has an associated n-dimensional reference vector of weights, one for each element of the input vector. In an abstract sense, the SOM functions compare the distance, in some suitable form, between each input vector and each reference vector in an iterative manner. With each iteration, the reference vectors are moved around in the output space until their positions converge to a stable state. When the reference vector that is closest to a given input vector is found (the winning reference vector), the reference vector is updated to more closely match the input vector. This is just the learning step.

LVQ uses the same internal architecture as SOM: a set of n-dimensional input vectors is mapped onto a two-dimensional lattice, and each node on the lattice has an associated n-dimensional reference vector. The learning algorithm in LVQ, i.e., the method of updating the reference vectors, is different from that in SOM. Because LVQ is a supervised method, during the learning phase the input data are tagged with their correct class and each output neuron represents a known category. We define the input vector $\boldsymbol{x}$ as

$$\boldsymbol{x} = (x_1, x_2, x_3, \cdots, x_n)\,,$$

and the reference vector for ith output neuron ω_i as

$$\boldsymbol{\omega}_i = (\omega_{1i}, \omega_{2i}, \omega_{3i}, \cdots, \omega_{ni})\,.$$

Define Euclidean distance between the input vector and the reference vector of the i neuron as

$$D(i) = \sqrt{\sum_{j=1}^{n}(x_j - \omega_{ji})^2}\,.$$

When $D(i)$ is a minimum, the input vectors are compared to the reference vectors and the closest match is found. The winning reference vector, ω_{i^*} is then obtained by the formula

$$|\ \boldsymbol{\omega}_{i^*} - \boldsymbol{x}\ | \leq |\ \boldsymbol{\omega}_i - \boldsymbol{x}\ |\ .$$

The reference vectors are then updated using the following rules:

$$\boldsymbol{\omega}_{i^*}(\text{new}) = \boldsymbol{\omega}_{i^*}(\text{old}) + \alpha(t)(\boldsymbol{x} - \omega_{i^*}(\text{old})) \qquad \text{if } \boldsymbol{x} \text{ is in the same class as } \boldsymbol{\omega}_{i^*},$$

$$\boldsymbol{\omega}_{i^*}(\text{new}) = \boldsymbol{\omega}_{i^*}(\text{old}) - \alpha(t)(\boldsymbol{x} - \boldsymbol{\omega}_{i^*}(\text{old})) \qquad \text{if } \boldsymbol{x} \text{ is in a different class from } \boldsymbol{\omega}_{i^*},$$

$$\boldsymbol{\omega}_{i^*}(\text{new}) = \boldsymbol{\omega}_{i^*}(\text{old}) \qquad \text{if } i \text{ is not the index of the winning reference vector.}$$

The learning rate $0 < \alpha(t) < 1$ should generally be made to decrease monotonically with time, with large changes in early iterations and more fine tuning as convergence is approached. There are several versions of the LVQ algorithm for which the learning rules differ in some details, see Kohonen (1995) for an explanation of the differences between these algorithms. When the learning phase is over, the reference vectors can be frozen, and any further inputs to the system will be placed into one of the existing classes, but the classes will not change.

3 CHOSEN SAMPLE AND PARAMETERS

The ROSAT Bright Source (RASS/BSC; Voges et al. 1999) contains positions, X-ray count rates, and spectral information of 18 811 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 (RASS/FSC) includes 105 924 sources. A catalogue of quasars and active nuclei (Véron-Cetty & Véron, 2000) which is an updated version of the catalogue of quasars and active nuclei (Véron-Cetty & Véron, 1998), contains 13 214 quasars, 462 BL Lac objects and 4428 active galaxies (of which 1711 are Seyfert 1s).

We perform positional cross-correlation of the catalogue of quasars and active nuclei with the ROSAT Bright Source Catalog (RASS/BSC) and Faint Source Catalog (RASS/FSC) X-ray sources in a search radius of 3 times their positional error, and then cross-identify the result with optical sources in the USNO A-2.0 catalog within 5 arcsec radius. After crossing out the one-to-many sources, the number of quasars, BL Lac objects and active galaxies reduces to 2272, 336, 1483, respectively. Similarly, using these sources to positionally cross-match 2MASS released data within 10 arcsec radius, after crossing out the one-to-many sources we obtained 909 quasars, 135 BL Lacs and 612 active galaxies. Likewise, we took stars from SIMBAD database and galaxies from Third Reference Catalogue of Bright Galaxies (RC3; de Vaucouleurs et al. 1991) to obtain a dataset of 9967 stars and 484 normal galaxies from optical and X-ray bands, and 3718 stars and 173 normal galaxies from optical, X-ray and infrared bands. For clarity, the adopted samples and corresponding catalogs are listed in Table 1. The chosen parameter, definition, catalogue and waveband are summarized in Table 2. The chosen parameters from different bands to classify objects are $B-R$ (optical index), $B+2.5\log$ (ct), ct (source count-rate in the broad energy band), HR1 (hardness ratio 1), HR2 (hardness ratio 2), ext (source extent), extl (likelihood of the source extent), $J-H$ (infrared index), $H-K$ (infrared index), $J+2.5\log$ (ct). The mean values of the parameters for all types of objects are presented in Table 3.

Table 1 Sample and Catalogue

Class of objects	Sample Size	Catalogue
Quasars	909	1^a
BL Lacs	135	1^a
Active galaxies	612	1^a
Stars	3718	2^b
Normal galaxies	173	3^c

1^a is the catalogue of quasars and active nuclei (Véron-Cetty & Véron, 2000).
2^b is SIMBAD database.
3^c is Third Reference Catalogue of Bright Galaxies (RC3; de Vaucouleurs et al. 1991).

Table 2 Chosen Parameter Summary

Parameter	Definition	Catalogue	Waveband
B	blue magnitude	USNO-A2.0	optical band
R	red magnitude	USNO-A2.0	optical band
ct	source countrate in the broad energy band	RASS/BSC,RASS/FSC	X-ray band
HR1	hardness ratio 1 Definition: hr1 = (B–A)/(B+A), where A=countrate in PHA range 11–41 B=countrate in PHA range 52–201	RASS/BSC,RASS/FSC	X-ray band
HR2	hardness ratio 2 Definition: hr2=(D–C)/(D+C), where C=countrate in PHA range 52–90 D=countrate in PHA range 91–201	RASS/BSC,RASS/FSC	X-ray band
ext	source extent	RASS/BSC,RASS/FSC	X-ray band
extl	likelihood of source extent	RASS/BSC,RASS/FSC	X-ray band
J	J band photometry	2MASS	infrared band
H	H band photometry	2MASS	infrared band
K	K band photometry	2MASS	infrared band

Table 3 Mean Values of Parameters for the Sample

Parameters	Quasars	BL Lacs	Active galaxies	Stars	Galaxies
$B-R$	0.11±0.51	0.78±0.91	0.78 ±0.89	–1.53±4.19	1.42±1.49
$B+2.5\log$ (ct)	13.87±1.09	15.18±1.57	13.02±2.40	4.18±5.33	7.95±2.40
ct	0.13±0.30	0.45±0.75	0.25±0.47	0.12±0.42	0.08±0.13
HR1	0.03±0.54	0.23±0.46	0.16±0.51	0.09±0.53	0.65±0.37
HR2	0.14±0.45	0.17±0.32	0.14±0.36	–0.02±0.54	0.22±0.48
ext	5.06±8.80	10.08±11.61	7.26±9.68	4.21±9.72	16.11±32.12
extl	1.15±4.49	5.38±15.23	2.20±5.83	1.05±6.74	7.81±31.15
$J-H$	0.68±0.27	0.75±0.14	0.79±0.15	0.02±15.97	0.76±0.17
$H-K$	0.79±0.31	0.70±0.17	0.75±0.23	–1.22±21.34	0.37±0.19
$J+2.5\log$ (ct)	12.87±0.96	13.54±1.42	12.54±1.53	–7.30±33.34	9.75±1.54

As shown by Table 3, different objects show different properties in different bands. Therefore it is reasonable to classify the objects with these parameters. To determine the best combination of parameters to discriminate between AGNs, stars and galaxies, we have probed a ten-dimensional space. With a principal component analysis, we found that a simple or weighted combination of the attributes forming a "super" attribute is not optimal. As a result, we may apply learning vector quantization (LVQ) more effectively and combine all or at least most of the attributes.

4 RESULTS

Taking the sample from optical, X-ray and infrared bands as the training set and the test set, we classify the sample into five classes by means of LVQ. Table 4 summarizes the classified result. The fractions of correct classification of quasars, BL Lacs, active galaxies, stars and normal galaxies are 84.2%, 31.1%, 60.8%, 97.3%, 65.9%, respectively. For the whole sample, the fraction is 88.5%. The results for quasars and stars are better than for the rest. Comparatively, the accuracies for BL Lacs, active galaxies and normal galaxies are low. In optical, X-ray and infrared bands, BL Lacs are hard to separate from quasars and active galaxies, simultaneously, active galaxies are not obviously different from quasars, while normal galaxies evidently differ from stars in that they are redder in the infrared band. Let "active objects" include quasars, BL Lacs and active galaxies and "non-active objects" include stars and normal galaxies. Then, as Table 4 shows, among 1656 active objects 40 or 2.4% are misclassified as non-active objects; among 3890 non-active objects, 116 or 3.0% are misclassified as active objects. Obviously, active objects can be separated from non-active objects in the three bands.

Table 4 Classified Result for the Multi-class Problem

	Known				
Classified	Quasars	BL Lacs	Active galaxies	Stars	Galaxies
Quasars	765	48	186	25	1
BL Lacs	20	42	19	10	2
Active galaxies	121	44	372	40	38
Stars	3	0	13	3616	18
Galaxies	1	1	22	27	114
Accuracy	84.2%	31.1%	60.8%	97.3%	65.9%

Since Table 4 shows that active objects apparently differ from non-active objects, we divide the sample into two parts: 1656 active objects and 3891 non-active objects. Then we apply LVQ to classify them. The sample acts as both the training set and the test set. The result is given in Table 5. The final accuracy is 97.6% for active objects and 97.1% for non-active objects. For the whole sample, the accuracy is 97.3%.

Table 5 Classified Result for the Two-Class Problem

	Known	
Classified	Active sources	Non-active sources
Active sources	1617	112
Non-active sources	39	3778
Accuracy	97.6%	97.1%

Now, we divide the sample in two parts, one of which is used as the training set, the other, the test set. The result of classification with LVQ is shown in Table 6. The accuracy of active objects is 96.0% and that of non-active objects is 96.1%. Among 2773 of the test sets, the number of correct classification is 2664, or 96.1%, the number of misclassifications is 109 or 3.9%. In order to further validate the effectiveness of LVQ, we used the data of Wei et al. (1999) and Xu et al. (2001) as test sets. By positional cross-correlation with the catalogues USNO-A2.0, RASS/BSC, RASS/FSC and 2MASS, we obtained 15 Seyfert galaxies, 12 quasars and two stars. By means of the LVQ, one Seyfert galaxy and two stars were misclassified. So the accuracy is 89.7%.

Table 6 Classified Result Using the Separated Sample for the Two-class Problem

	Known	
Classified	Aactive sources	Non-active sources
Active sources	795	76
Non-active sources	33	1869
Accuracy	96.0%	96.1%

5 DISCUSSION

The methods of selecting quasar candidate employed by previous surveys include selections by radio, color, slitless spectroscopy, X-ray and infrared sources, variability, or zero proper motion. However, the main drawback of radio selection is that most, if not all, radio-quiet quasars cannot be included in the sample. So the sample cannot be representative of the quasar population as a whole. The color selection and slitless spectroscopy selection are efficient, but they both suffer significant selection effects. Now, strong X-ray emission has been found to be nearly the defining characteristic of AGNs. Therefore X-ray selection from a very deep X-ray survey might be the best way to obtain the most complete census of AGNs. However, because soft X-rays can be easily absorbed by the intergalactic medium, distant or high-redshift AGNs are always weak in X-ray emission and may not be detected by present X-ray telescopes. Existing X-ray surveys also show that X-ray selection is heavily biased against high-redshift

objects. The last three methods are not effective in practice and few surveys have been based on them. Thus, all the ways described here suffer significant selection effects, and the resulting AGN samples are all incomplete to some extent. In order to construct samples with a high degree of completeness, combined methods need to be employed. Wei (1999) used $\log C + 0.4R$ as an alternative expression for $\log(f_{\rm X}/f_{\rm opt})$, where C is X-ray count rate and R stands for the R magnitude. Then a criterion of high X-ray-to-optical flux ratio, i.e. $\log C \geq -0.4R + 4.9$, was used to pre-select AGN samples. The success rate of detecting AGN then amounted to 73%. This method only selects AGNs with high X-to-optical flux ratio, but cannot acquire complete samples. In this letter, we obtain the sample from optical, X-ray and infrared bands, and put forward an automatic method, learning vector quantization (LVQ), to classify multiwavelengh data. Compared to other methods, our method avoids the weakness of a single selection method, artificial cutoffs, and improves the accuracy and efficiency of pre-selecting sources.

At present, we cannot effectively categorize by LVQ the sample to five types: quasars, BL Lacs, active galaxies, stars and galaxies with the ten parameters. The low accuracy of BL Lacs, active galaxies and galaxies possibly results from their small sample sizes, and also from the fact that the combination of the parameters is not ideal for discriminating between AGNs and non-AGNs. To improve the accuracy, we need to enlarge the size of the sample or to extract more effective features. In fact, if we analyze the classification, we may note that contamination is probably due to two reasons: (1) true misclassifications and (2) a priori classification selection problem. The former is inherent for the chosen method. The latter refers to the fact of having a restricted number of output classes, for instance, when a star-galaxy classifier is presented with cases of processing defects, source merging, or cosmic rays. Hence, one must ensure that in the case of automatic classification there should be enough classification categories available, otherwise the data must be preprocessed correspondingly. The recipe for more successful neural network classification could be to provide enough neurons to learn the categories. Overall the method is potentially very promising because they can provide a more robust approach to source segmentation.

6 CONCLUSIONS

In this paper we have introduced the learning vector quantization algorithm (LVQ) applied to the data from optical, X-ray and infrared bands, and tested it with different samples. LVQ shows better performance in classifying multi-wavelength data, especially in separating AGNs from stars and normal galaxies. Considering classification reliability, we find that the parameters from optical, X-ray and infrared bands that lead to the most reliable classification results are $B - R$, $B + 2.5\log$ (ct), ct, HR1, HR2, ext, extl, $J - H$, $H - K$ and $J + 2.5\log$ (ct), they are proved to be effective features for separating AGNs from non-AGNs. Our results will be applicable to preselect sources in other digital and digitized sky surveys that will be a part of International Virtual Observatory (IVO). A careful, systematic selection of targets from large sky surveys for focused follow-up studies using large telescopes and space observatories would make an optimal use of the valuable observing time at such costly facilities. Perhaps the most important would be the enabling role of IVO in making these information-rich data sets and tools to explore them available to the broad community, regardless of their access to large telescopes: important new discoveries can be made in data mining of the digital sky. With the data improving in quality and quantity, LVQ will show its superiority in classification in astronomy. One possible extension of this work is to add parameters from more bands, or

change the extracted features, for instance, to select features from spectral information of all kinds of objects. Another can be the application of unsupervised methods to these data sets, which could result in a different, new classification scheme.

Acknowledgements We gratefully acknowledge the help of the LAMOST staff, in particular Dr. Cui Chenzhou for source identifications, Dr. Luo Ali, Wang Wei and Dr. Bian Weihao for helpful discussion. We are also very grateful to the referee's significant comments. This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France. 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 the Space Administration and the National Science Foundation, USA. This project is supported by the National Natural Science Foundation of China under grant 10273011.

References

Allende Prieto C., Rebolo R., Lopez R. J. G. et al., 2000, AJ, 120, 1516
Andreon S., Gargiulo G., Longo G. et al., 2000, MNRAS, 319, 700
Bailer-Jones C. A. L., Irwin M., von Hippel T., 1998, MNRAS, 298, 361
Bazell D., Peng Y., 1998, ApJS, 116, 47
Bertin E., Arnout S., 1996, AAS, 117, 393
Bishop C. M., 1995, Neural Networks for Pattern Recognition, Oxford University Press
de Vaucouleurs G., de Vaucouleurs A., Corwin H.G. et al., 1991, Third Reference Catalogue of Bright Galaxies (RC3), New York: Springer-Verlag
Kohonen T., 1989, Self-Organization and Associative Memory (3d ed.), Berlin: Springer
Kohonen T., 1990, Proc. IEEE, 78, 1464
Kohonen T., 1995, Self-Organization Maps, Berlin: Springer
Lahav O., Naim A., Sodre L. Jr. et al., 1996, MNRAS, 383, 207
Mähönen P. H., Hakala P. J., 1995, ApJ, 452, L77
Naim A., Lahav O., Sodre L. Jr. et al., 1995, MNRAS, 275, 567
Odewahn S. C., Stockwell E. B., Pennington R. L. et al., 1992, AJ, 103, 318
Storrie-Lombardi M. C., Lahav O., Sodre L. Jr. et al., 1992, MNRAS, 259, 8
Véron-Cetty M. P., Véron P., 1998, ESO Scientific Report 18
Véron-Cetty M. P., Véron, P., 2000, ESO Scientific Report 19
Voges W., Aschenbach B., Boller Th. et al., 1999, A&A, 349, 389
Weaver W. B., 2000, ApJ, 541, 298
Wei J. Y., Xu D. W., Dong X. Y. et al., 1999, A&AS, 139, 575
Xu D. W., Wei J. Y., Hu J. Y., 2001, Chin. J. Astron. Astrophys. (ChJAA), 1, 46