P.068

LS-SVM Applied to Photometric Classification of Quasars and Stars

Yanxia Zhang, Yongheng Zhao and Nanbo Peng,
Key Laboratory of Optical Astronomy, National Astronomical Observatories;
Chinese Academy of Sciences, 100012 Beijing, China

Abstract

The major demerit of Support vector machines (SVM) is its higher computational cost for a quadratic programming (QP) problem. In order to overcome this problem, Least Squares Support Vector Machines (LS-SVM) is put forward. LS-SVM's solution is given by a linear system, which makes SVM method more generally simple and applicable. In this paper, LS-SVM is used for classification of quasars and stars from SDSS and UKIDSS photometric databases. The result shows that LS-SVM is highly efficient and powerful especially for large scale problem and has comparable performance with that of SVM.

Introduction

With the construction of the space- and ground-based telescopes and the advancement of detection and reception technologies, astronomy enters an era abundant in data and information. Astronomical data increase at a breathtaking speed, up to Petabyte. How to collect, save, analyze and mine so huge data is a challenge for astronomers. Astronomers have to fall back on the uprising and flourishing data mining technology. Most of astronomical problems belong to the different tasks of data mining, and can be solved by data mining methods. The detailed review about this issue can refer to the review papers (Zhang, et al. 2002, Zhang, et al. 2008, Ball 2010)

Quasars and stars all belong to pointed sources from images, so successful selection of quasars from stars is of great importance. Quasar candidate preselection methods applied 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. All of these methods have merits and demerits. In order to improve the effectiveness and efficiency of large sky survey, careful preparation of input catalogue is necessary. With large photometric data, the high-efficient rapid classification algorithms are in need. This paper focuses on studying the capabilities of LS-SVM to derive accurate and robust classification models for the type prediction of the unclassified photometric data from large sky survey data. The performance of LS-SVM classification will be comparable to SVM in terms of accuracy and be superior to SVM on running time.

Data

We adopted the same sample as that is studied in Wu & Jia (2010).The quasar sample is obtained by cross-identification of all quasars in SDSS Data Release 7 (DR7) with the UKIDSS Data Release 3 (DR3)within 3 arcsecond radius. The star sample is similarly obtained by cross-match of the two survey databases. The final sample includes 8498 quasars and 8996 stars with both SDSS ugriz and UKIDSS YJHK.

Method

Least-Squared Support Vector Machine (LS-SVM), a semi-parametric modeling technique, is the least squares version of Support Vector Machine (SVM). The main reference and overview on LS-SVM is detailed in Suykens et al. (2002). We apply the toolbox LS-SVM in Matlab implementation. The software can be downloaded from the website: http://www.esat.kuleuven.be/sista/lssvmlab/.

Results

We tried many experiments with the above sample by LS-SVM and compared the performance based on different input patterns and different model parameters. In order to determine the useful input pattern, we randomly separate the sample into three parts: two thirds for training and one third for testing. The result is shown in Table 1. The model parameters and 2 may be determined by the program itself. Considering the running time, we randomly tested some and 2 values, and kept the better result. When the model parameters are set, the time to construct a predicting model is very short, only costing one or more minutes. The used computer is Intel(R) Core(TM)2 Quad CPU Q9550@2.83GHz with memory of 6GB. As Table 1 shows, the performance based on (*ugrizYJH K*) is superior to that on (*ugriz*) or (*YJHK*); that of (*YJHK*) is better than that of (*ugriz*). Moreover the accuracy of (*u-g,g-r,r-I,i-z,z-Y,Y-J,J-H,H-K*) is higher than that of (*u-g,g-r,r-I,i-z*) or (*Y-J,J-H,H-K*); that of (*Y-J,J-H,H-K*) outperforms that of (*u-g, g-r,r-I,i-z*). The best performance adds up to 97.98% with input pattern of (*u-g,g-r,r-I,i-z,z-Y,Y-J,J-H,H-K*) and model parameters (gamma= 10 and Sigma^2 = 4); the better one is 97.38% with (*ugrizYJHK*) and the same model parameters. Since the result with colors shows superiority to that with magnitudes, we further experiment the sample with colors by three-fold cross-validation. The result is indicated in Table 2. The accuracy with 8 colors amounts to (98.81+_0.60)%; that with 3 infrared colors is (95.36+_1.97)%; that with 4 optical colors is only (82.37+_7.21)%. Here we only discuss the result by three-fold cross-validation, not by ten-fold cross-validation because the latter needs much larger computer memory which surpasses the present memory of my computer. For a large enough sample, the three-fold cross-validation is reliable and applicable.

Table 1. The performance of different input patterns

input pattern/model parameter	accuracy $\gamma = 10, \sigma^2 = 0.4$	accuracy $\gamma = 10, \sigma^2 = 4$
$u, g, r, i, z, Y, J, H, K$	96.69%	97.38%
u, g, r, i, z	91.96%	90.67%
Y, J, H, K	92.23%	92.03%
$u-g, g-r, r-i, i-z, z-Y, Y-J, J-H, H-K$	95.94%	97.98%
$u-g, g-r, r-i, i-z$	92.99%	92.39%
$Y-J, J-H, H-K$	94.48%	94.41%

Table 2. The classification accuracy of three-fold cross-validation

	average accuracy
$u-g, g-r, r-i, i-z, z-Y, Y-J, J-H, H-K$	$(98.81 \pm 0.60)\%$
$u-g, g-r, r-i, i-z$	$(82.37 \pm 7.21)\%$
$Y-J, J-H, H-K$	$(95.36 \pm 1.97)\%$

CONCLUSIONS

We explored LS-SVM to classify quasars from stars with SDSS DR7 and UKIDSS DR3 databases. Efficient methods to preselect quasar candidates are of great value for large sky survey projects, such asthe Chinese Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST, now renaming Guoshoujing Telescope). In our case, LS-SVM shows its advantage of short running time and good performance. Nevertheless this approach is memory consuming. Therefore for the much larger sample, LS-SVM shows its weakness. In future work, we further study how to accelerate the speed of data mining techniques and keep good performance at the same time.

CONTACT

E-mail:
zyx@bao.ac.cn;
yzhao@lamost.org;
nbpeng@bao.ac.cn;
Phone: 8610-64841693
Fax: 8610-64878240

Poster Design & Printing by Genigraphics® - 800.790.4001