APPLYING THE SUPPORT VECTOR MACHINE METHOD TO MATCHING *IRAS* AND SDSS CATALOGUES

Chen Cao
National Astronomical Observatories, CAS, Beijing 100012, China
Graduate School, Chinese Academy of Sciences, Beijing 100039, China
Email: caochen@bao.ac.cn

ABSTRACT

This paper presents results of applying a machine learning technique, the Support Vector Machine (SVM), to the astronomical problem of matching the Infra-Red Astronomical Satellite (IRAS) and Sloan Digital Sky Survey (SDSS) object catalogues. In this study, the IRAS catalogue has much larger positional uncertainties than those of the SDSS. A model was constructed by applying the supervised learning algorithm (SVM) to a set of training data. Validation of the model shows a good identification performance (~ 90% correct), better than that derived from classical cross-matching algorithms, such as the likelihood-ratio method used in previous studies.

Keywords: Miscellaneous astronomical data bases, Catalogs, Surveys, Sloan Digital Sky, Infra-red astronomy

1 INTRODUCTION

Matching catalogues of multi-wavelength observations is one of the fundamental problems in astronomical research (e.g. Cao et al., 2006; Rohde et al., 2005, 2006), especially for analyzing the vast amount of data being collected by sky surveys (e.g., SDSS). Simple matching approaches such as the 'closest match,' which depend on positions only, are considered adequate only when the positional uncertainties of the matched catalogues are all very small (e.g., matching SDSS and *Spitzer* IRAC catalogues, Wu et al., 2005). Another technique, the likelihood-ratio method (Sutherland & Saunders, 1992), which compares the probability that the object is a 'true' match with the probability that it is a chance background source, can be used for matching catalogues with large positional uncertainties (e.g., matching SDSS and IRAS catalogues, Cao et al., 2006). However, this technique can only utilize a small number of the catalogues' parameters (positions and magnitudes) and is also model dependent.

Supervised learning techniques from machine learning have been widely used for classification and regression, and recently they have been extended to the domain of catalogue matching problems. Rohde et al. (2005) applied automated machine learning techniques to the problem of matching catalogues where one of the two has large positional uncertainties. In their study, a model was first constructed based on a supervised learning algorithm and a set of training data. Validation of the model shows a very good (>99% correct) identification performance. In this work I apply a similar matching technique, the Support Vector Machine, to matching SDSS and *IRAS* catalogues. The performance is validated using two-fold cross-validation, and the results are better than those derived from the likelihood ratio method used in previous studies (Cao et al., 2006).

2 CATALOGUE DESCRIPTIONS

2.1 *IRAS* Faint Source Catalogue

The Infra-Red Astronomical Satellite (*IRAS*) was launched in 1983 (Neugebauer et al., 1984) and scans almost all the sky in mid- and far-infrared (12, 25, 60, 100 μm) wavebands. The Faint Source Catalogue (FSC, Version 2.0, Moshir+ 1989) was released after the Point Source Catalogue (PSC, IPAC 1986). It contains data for 173,044 point sources in unconfused regions with flux densities typically above 0.2 Jy at 12, 25 and 60 μm and above 1.0 Jy at 100 μm and achieves roughly one-magnitude deeper in sensitivity relative to the PSC. The catalogues give the *IRAS* sources' four band flux densities and qualities, the positions of the sources, and other useful parameters. The sources in the catalogues all have large positional uncertainties, which can be described as an "error ellipse." The error ellipse stands for the uncertainties along (in-scan) and across (cross-scan) the *IRAS*'s scan direction, and the uncertainty ellipse major axis, minor axis, and positional angle in the catalogues are used for describing it.

2.2 SDSS Catalogue

The Sloan Digital Sky Survey (SDSS, York et al., 2000) contains an imaging survey of the northern sky in the five bands u, g, r, i, z and a spectroscopic target survey performed by multi fibers. The Second Data Release (DR2, Abazajian et al., 2004) was released in 2004. The SDSS-DR2 spectroscopic target survey covers about 2627 deg^2 of the sky, including about 260,490 galaxies, 32,241 quasars, 3,791 high-z ($z > 2.3$) quasars, and others objects. For the study of the detailed spectral properties of infrared galaxies (such as their emission lines), we only choose the SDSS-DR2 spectroscopic targets with a redshift greater than 0.001 (to reject stars) and high redshift confidence ($zConf > 0.9$) to do the cross-matching with *IRAS*.

3 APPLYING THE SVM METHOD TO MATCHING *IRAS* AND SDSS CATALOGUES

3.1 Matching *IRAS* and SDSS Catalogues

First we use the *IRAS* error ellipse as the cross-section (the SDSS's position uncertainties are neglected compared with the *IRAS*'s) to do cross-matching with the SDSS sources spectral positions. Two RMS uncertainty (2σ) significance levels were chosen for a high level of confidence and more complete sample selection. However, it is not easy to determine whether the matched SDSS targets are really the infrared objects. In a previous study (Cao et al., 2006) we used the "Likelihood Ratio" (LR) method (Sutherland & Saunders, 1992) to calculate the probability of the "true" cross-correlation for each matched SDSS object. This method can only utilize a small number of parameters in the catalogues (e.g., positions and magnitudes) and is also limited by many assumptions (e.g., the errors should be a Gaussian distribution). Thus, the Support Vector Machine method could be used here to solve these problems.

3.2 Support Vector Machine Method

The foundations of Support Vector Machines (SVM) have been developed by Vapnik (1995). They are promising methods for data classifications and regressions and have been successfully applied to many

astronomical problems (e.g., Zhang et al., 2003). In the support vector classification problem, the goal is to find an optimal separating hyperplane, which can maximize the distance (margin) between it and the nearest data point of each of the two classes. The success of SVMs in practice mainly comes from their properties of handling a nonlinear classification/regression efficiently using kernel functions, which can transform the input space into another high-dimensional feature space.

3.3 Results and Performances

The SVM software being used here is SVM^{light} (Joachims, 1998), with popular kernel functions including Linear, Polynomial (d=2,3), and Radial Basis Functions (RBF, γ=1). Soft margin is also used for the SVM to avoid overfitting (Cristianini & Shawe-Taylor, 2000). The choice of features in this study are given in Table 1: the separation of targets from SDSS and *IRAS* catalogues (in units of σ); the SDSS five-band (u,g,r,i,z) magnitudes; and the *IRAS* four-band (12, 25, 60, 100μm) flux densities. The performances of the identification are shown in Table 2. Two-fold cross-validations are used here based on a set of training data (selected based on previous works) and a set of testing data, as shown in Figure 1. The performance (~ 90% correct; Polynomial, d=2) is better than that derived from the likelihood ratio method used in previous studies (Cao et al., 2006).

Feature	Origin	Name
1	SDSS-*IRAS*	Separation (σ)
2	SDSS	mag_u
3	SDSS	mag_g
4	SDSS	mag_r
5	SDSS	mag_i
6	SDSS	mag_z
7	*IRAS*	f12 (Jy)
8	*IRAS*	f25 (Jy)
9	*IRAS*	f60 (Jy)
10	*IRAS*	f100 (Jy)

Table 1. Features for SVM inputs

Algorithm (Kernel)	Soft margin (c)/hu	Accuracy 2-fold
Linear	0.1	70.89%
	1	71.29%
	10	70.81%
Poly (d=2)	0.1	90.00%
	1	89.69%
	10	88.67%
Poly (d=3)	0.1	44.28%
	1	44.28%
	10	44.28%
RBF (γ=1)	0.1	81.85%
	1	87.57%
	10	86.22%

Table 2. Performance of the algorithms and parameters

.

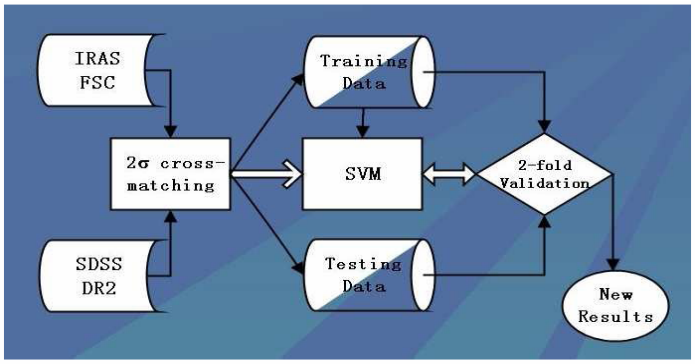

Figure 1. Flowchart of applying SVM technique to matching *IRAS* and SDSS catalogues

4 DISCUSSION

Matching catalogues with significant positional uncertainties are typical problems in the Virtual Observatory (VO), which will (and already have brought) bring new challenges and opportunities to modern astronomical research. Using supervised learning techniques such as the SVM method for matching catalogues in VOs is a big challenge and is more difficult than the simple matching algorithms (tools) used in current VO systems (e.g., the 'Matcher' software in the GAVO). However, these techniques could become a good solution to many matching problems in VO systems that cannot be dealt with by classical cross-matching algorithms.

5 ACKNOWLEDGEMENTS

I would like to thank Drs. H. Wu, D. Rohde, Y.-X. Zhang, C. Liu, D. Gao, and D. Wang for advice and helpful discussions. The SDSS Web site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions.

6 REFERENCES

Abazajian, K., et al. (2004) *AJ 128*: 502.

Cao, C., Wu, H., Wang, J.-L., Hao, C.-N., Deng, Z.-G., Xia, X.-Y., & Zou, Z.-L. (2006) *Chinese Journal of Astronony and Astrophysics 6:* 197.

Cristianini N. & Shawe-Taylor J. (2000) *Support Vector Machines*. Cambridge Univ. Press: Cambridge.

Joachims T. (1998) in Schölkopf B., Burges C., & Smola A. (eds.) *Advances in Kernel Methods: Support Vector Machines. MIT* Press: Cambridge, MA.

Moshir, M., et al. (1990) IRAS Faint Source Catalogue, version 2.0.

Rohde, D. J., Drinkwater, M. J., Gallagher, M. R., Downs, T., & Doyle, M. T. (2005) *MNRAS 360*: 69.

Rohde, D. J., Gallagher, M. R., Drinkwater, M. J., & Pimbblet, K. A. (2006) *MNRAS 369: 2.*

Sutherland, W. & Saunders, W. (1992) MNRAS 259: 413.

Vapnik V. (1995) *The Nature of Statistical Learning Theory.* Springer: Berlin.

Wu, H., Cao, C., Hao, C.-N., Liu, F.-S., Wang, J.-L., Xia, X.-Y., Deng, Z.-G., & Young, C. K.-S. (2005) *ApJL, 63:* L79.

York, D. G., et al. (2000) *AJ 120:* 1579.

Zhang, Y. & Zhao, Y. (2003) *PASP 115*: 1006.