Automatised classification of WISE sources: first results, future prospects

Agnieszka Kurcz[1·2], Magdalena Krupa[1·2], Maciej Bilicki[3·2·4], Aleksandra Solarz[5·2], Agnieszka Pollo[1·5·2] and Katarzyna Małek[5·2]

1. Astronomical Observatory, Jagiellonian University, ul. Orla 171, 30-244 Kraków, Poland
2. Janusz Gil Institute of Astronomy, University of Zielona Góra, ul. Lubuska 2, 65-265 Zielona Góra, Poland
3. Leiden Observatory, Leiden University, P.O. Box 9513 NL-2300 RA Leiden, The Netherlands
4. Astrophysics, Cosmology and Gravity Centre, Department of Astronomy, University of Cape Town, Rondebosch, South Africa
5. National Centre for Nuclear Research, Astrophysics Division, ul. Hoża 69, 00-681 Warszawa, Poland

We present the first results of our dedicated programme of automatised classification of galaxies, stars and quasars in the mid-infrared all-sky data from the WISE survey. We employ the Support Vector Machines (SVM) algorithm, which defines a hyperplane separating different classes of sources in a multidimensional space of arbitrarily chosen parameters. This approach consists of four general steps: 1) selection of the training sample, 2) selection of the optimal parameter space, 3) training of the classifier, 4) application to target data. Here, as the training set, we use sources from a cross-correlation of the WISE catalogue with the SDSS spectroscopic sample. The performance of the SVM classifier was tested as a function of size of the training set, dimension of the parameter space, WISE apparent magnitude and Galactic extinction. We find that our classifier provides promising results already for three classification parameters: magnitude, colour and differential aperture magnitude. Completeness and purity levels as high as 95% are obtained for quasars, while for galaxies and stars they vary between 80–95% depending on the magnitude, deteriorating for fainter sources.

1 Introduction

Today's flood of astronomical data gathered in ever wider and deeper datasets comes at a price of inability to obtain spectra for the majority of the observed sources. Without knowing the spectral features, classification of objects in large photometric samples becomes far from straight-forward, especially at the faint end and in the low signal-to-noise regime. A good example is the all-sky catalogue from the Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010), providing various pieces of photometric and astrometric information for almost 10^9 sources, without however any object type identification. For such amount of data, human visual verification of sources is clearly infeasible except for some very small subsamples. Limited number of photometric bands and a low detection rate in some of them, together with sources overlapping in multi-colour space, make also traditional approaches to separate objects, such as through colour cuts, not always effective in this dataset.

Current approaches towards identifying specific source types in WISE consist mainly in cross-matching this dataset with an external one (e.g. SDSS) and calibrating some magnitude and colour cuts to be applied for a specific subsample preselection (e.g. Stern et al. 2012 for AGNs or Tu & Wang 2013 for AGB stars). Application of such cuts to the all-sky WISE data may however give biased results, for such reasons as non-representativeness of the calibration sample, variations in WISE source detection rate (cf. Secrest et al. 2015) or blending and varying stellar populations depending on sky position, leading to variations is source effective colours (e.g. Ferraro et al. 2015).

A possible way to avoid these issues is to rely on automatised classification through machine learning (ML) algorithms, such as Support Vector Machines (SVM). An example application using WISE data is provided in Kovács & Szapudi (2015), where however the depth of the resulting galaxy catalogue is limited by much shallower 2MASS. Our aim is to go beyond such limitations and apply ML to the majority of WISE sources, ideally at its full depth. Małek et al. in this volume and Krakowski et al. (in prep.) describe an independent work where the same methodology is applied to cross-matched WISE×SuperCOSMOS data (Bilicki et al., 2016). In the present article we discuss the results of various tests of SVM applied to WISE-only data.

2 Data description

2.1 WISE survey and our preselection

The data used in this work come from WISE (Wright et al., 2010), a NASA satellite mission launched in December 2009. This 40-cm telescope, with a total $47' \times 47'$ field of view, scanned the entire sky in four infrared bands ($W1$ – $W4$) centred at 3.4, 4.6, 12 and 22 μm. WISE provides much better sensitivity than all the earlier infrared all-sky surveys (including IRAS, Neugebauer et al. 1984; 2MASS, Skrutskie et al. 2006; and AKARI, Murakami et al. 2007), and is free of atmospheric contamination. This directly translates to much larger photometric depth, which for WISE is about 3 mag better than for 2MASS.

The WISE catalogues are publicly available[1] and contain positional, photometric and detection quality information, as well as motion fit parameters. Here we use data from its second all-sky release, 'AllWISE' (Cutri et al., 2013), containing about 750 million sources, which makes it one of the largest existing astronomical catalogues.

The goal of our work is to classify as many WISE sources as possible, therefore our basic source selection was not restrictive. In particular, we required the sources to have at least 2σ detection only in the two shortest WISE bands ($W1$ and $W2$), discarding much shallower $W3$ and $W4$ channels. A basic cleanup to ensure reliability of the sources (removal of saturated objects and artifacts), together with an overall cut of $W1 < 17$ (Vega) for uniformity left us with over 606 million sources on the whole sky (see Bilicki et al. 2016 for a map and other details). Note that owing to the 6" resolution of WISE, severe blending arises in the Galactic Plane and Magellanic Clouds, which makes efficient source identification practically impossible in these areas. On the other hand, WISE is only minimally affected by Galactic extinction, which is an order of magnitude smaller in the mid-IR than in the optical.

[1] `http://irsa.ipac.caltech.edu/`

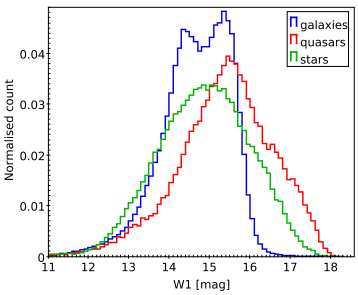

Fig. 1: Normalised $W1$ counts (Vega) for galaxies, quasars and stars in the WISE×SDSS DR10 spectroscopic sample.

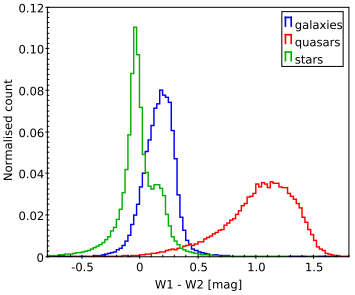

Fig. 2: Distribution of the $W1 - W2$ colour for galaxies, quasars and stars in the WISE×SDSS DR10 spectroscopic sample.

2.2 WISE×SDSS training sample

ML source classification, such as with SVM, relies on a training sample, containing sources of already identified types. In this work, as the training set we use the WISE data cross-matched with the spectroscopic sample from the SDSS Data Release 10 (Ahn et al., 2014). The resulting WISE×SDSS dataset contains 2.1 million sources in total, however to ensure its reliability, we cleaned it up of insecure measurements, using in particular the SDSS zWarning and zErr parameters. After imposing appropriate conditions on these quantities, we were left with about 390,000 stars, 1.5 million galaxies and 190,000 quasars in the training sample.

Figure 1 presents normalised $W1$ counts for the three source types in the WISE× SDSS cross-match. As is clearly visible, SDSS contains hardly any galaxies fainter than $W1 = 16$. For that reason, as we will be unable to create a reliable training sample beyond that magnitude, our analysis will be from now on restricted to $W1 < 16$. As far as all-sky data are concerned, this cut reduces the size of our WISE catalogue to 314 million sources. As $W1 = 16$ is one magnitude brighter than the overall completeness of WISE, our analysis will need to be extended when deeper training samples become available, such as for instance from SDSS-IV (Dawson et al., 2015).

We also note that the galaxy counts in WISE×SDSS are characterized by two distinct peaks. This results from the combination of two effects: heterogeneity of the SDSS spectroscopic data, and properties of WISE-detected galaxies seen also by SDSS. Finally, the training sample contains practically no galaxies nor quasars brighter than $W1 \sim 9.5$. This is however not a major issue in view of the eventual all-sky classification, as such WISE sources are mostly stars (Jarrett et al., 2011) and/or are saturated.

Figure 2 presents the distribution of the $W1 - W2$ colour for the three source types. For stars and galaxies, a significant overlap in this colour makes it insufficient as a separator in the absence of other source information, such as morphology (which is *not* provided in the WISE database). Even for quasars, which much more clearly separate out from other source types, the $W1 - W2 > 0.8$ cut proposed by Stern et al. (2012) is not fully appropriate, giving a potentially very incomplete sample. This further motivates our approach to identify WISE sources in multi-dimensional space in an automatised way rather than by simple cuts.

3 Methodology

3.1 SVM method of classification

Automatic classification of the WISE sources presented here uses a class of supervised ML algorithms – Support Vector Machines (SVMs). Similarly as some other ML classification methods, it relies on choosing an appropriate feature space, in which the sources from a given sample occupy different parts according to their class (here galaxies, stars and quasars). Thanks to an algorithm using pattern recognition – such as SVM – we can distinguish these classes in the multi-dimensional parameter space.

The main idea behind the SVM algorithm is to calculate a decision boundary between a set of different objects. Maximising the margin (i.e. the shortest distance from the decision plane to the closest points belonging to the distinct classes) between the classes closest points (the so-called support vectors), the optimal separating hyperplane between the N classes of sources can be found (in our case $N = 3$). For the experiments described here, we used the Gaussian radial basis kernel function, for which we tuned two parameters determining the separation boundary: C and γ. The parameter C is responsible for the width of the margin, while γ specifies the topology of the decision boundary. These parameters are fitted based on the training sample, and hence the decision surface can be established.

3.2 Efficiency of classification

For each of the cases described hereunder, two tests were made. In the first one (*self-check*), we classified the same objects as present in the training sample. In addition, for a *cross-test*, we applied the classifier to a randomly chosen sample of objects outside of the training set. To verify the efficiency of our method, we computed the completeness, C, purity, P and contamination, F. For galaxies they are given by (e.g. Soumagnac et al. 2015):

$$\mathrm{C_g} = \frac{\mathrm{TGG}}{\mathrm{TGG} + \mathrm{FGS} + \mathrm{FGQ}} \,, \tag{1}$$

$$\mathrm{P_g} = \frac{\mathrm{TGG}}{\mathrm{TGG} + \mathrm{FSG} + \mathrm{FQG}} \,, \tag{2}$$

$$\mathrm{F_g} = 1 - \mathrm{P_g} \,, \tag{3}$$

where TGG, FGS and FGQ refer to true galaxies classified respectively as galaxies, stars or quasars, and FSG, FQG are stars or quasars misclassified as galaxies. Similar statistics were computed for stars and quasars.

3.3 Details of tests performed on WISE data

For the classification tests presented here, we divided the WISE×SDSS sample according to WISE $W1$ magnitudes and Galactic extinction to examine the behaviour of the classifier as a function of these parameters. One expects the classification efficiency to deteriorate for fainter (hence lower signal-to-noise) sources on the one hand, while extragalactic sources located in sky areas of similar extinction should have their colours similarly biased, potentially influencing the results. We thus created three flux-limited subsamples of the training data ($W1 < 14$, < 15 and < 16), which

were further divided according to extinction. For the latter we used the I_{100} sky map, made from a combination of COBE/DIRBE and IRAS 100 μm measurements (Schlegel et al., 1998), and applied four bins: $I_{100} \in \langle 0;1), \langle 1;2), \langle 2;3)$ and $\langle 3;10)$ [MJy/sr]. The training set includes practically no galaxies above $I_{100} > 10$ MJy/sr; such areas cover however mostly the Galactic plane and Magellanic Clouds where classification is unreliable anyway.

In our experiments we have used the following quantities derived from the WISE database to define the parameter space:

1. magnitude w1mpro measured with profile-fitting photometry in the $W1$ band;
2. colour $W1 - W2$ defined as the difference in the w1mpro and w2mpro profile-fitting magnitudes;
3. difference of two circular aperture magnitudes in $W1$, w1mag_1 − w1mag_3, measured respectively in radii 5.5" and 11" ;
4. apparent motion defined as $\mathsf{pm} = \sqrt{\mathsf{pmra}^2 + \mathsf{pmdec}^2}$, where pmra and pmdec are the apparent motions in right ascension and declination, respectively.

We note that among several dozen of photometric quantities provided in the AllWISE database, only a small fraction have reliable measurements for all the WISE sources and are not strictly correlated with each other. For that reason, the parameter space potentially available for the all-sky classification is very limited.

4 Results

We verified how the classifier behaves as a function of: i) the size of the training set; ii) number of training parameters; iii) Galactic extinction; iv) limiting WISE magnitude. All these tests were done in the magnitude/extinction bins. Here we briefly describe the results. More details will be provided in Kurcz et al. (in prep.).

4.1 Size of the training set

We first verified what is the minimum size of the training set for which the SVM classifier produces stable results. For that purpose we used random subsamples of WISE×SDSS having 100, 1000, 3000 or 5000 objects of each class (i.e. 100 galaxies, 100 quasars and 100 stars, etc.). Figure 3 presents an example of results: completeness for a bin $W1 < 15$ for the cross-test. Classifier's performance stabilises for the samples with 9000 objects in total. Similar results were obtained for the other bins, we thus conducted the remaining tests for training sets of this size.

4.2 Dimension of the parameter space

Having established the optimal size of the training set, we performed a series of tests to verify how many parameters would suffice for reliable classification. In this study we limited ourselves to the 4 parameters defined in Sec. 3.3. The fourth of them, proper motions, is measured only for a fraction of WISE sources, so we used it mostly as a test-case for future, more precise measurements.

Results are presented in Figure 4, and the following parameter combinations are illustrated: magnitude $W1$ and colour $W1 - W2$ (2 parameters); the former with the

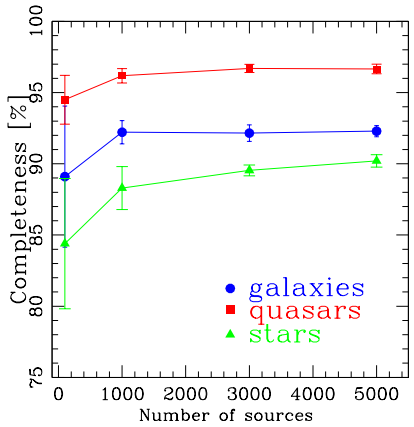

Fig. 3: Dependence of the completeness on the number of objects in the training set, for a bin of $W1 < 15$ and $I_{100} \in \langle 0; 1)$ for the cross-test case.

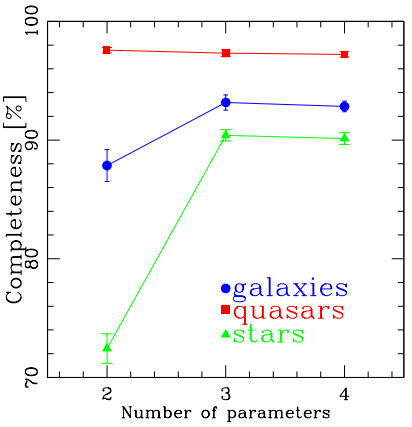

Fig. 4: Dependence of the completeness on the number of classification parameters, for a bin of $W1 < 15$ and $I_{100} \in \langle 0; 1)$ for the cross-test case.

differential aperture magnitude added (3 params.); with proper motions (4 params.). Adding the differential aperture magnitude (which serves as a morphological proxy) significantly improved the results, while including also the proper motions did not. We thus conducted the remaining tests for the 3-parameter case only.

4.3 Dependence of classification efficiency on extinction and magnitude

In the final series of tests, we examined the dependence of classification statistics on the apparent $W1$ magnitude and on Galactic extinction. Here the sampling of magnitudes was done in 0.5 mag bins, while the extinction bins remained the same as above. The results are summarised in Figures 5–7 for the 3 source classes. As far as the extinction is concerned, the differences between various bins are very small, consistent with no dependence of the performance on this parameter. The situation is different for magnitudes, where for galaxies and stars both the completeness and purity consistently decrease as the sources are getting fainter; there is however no such effect for quasars. Still, even at the faint end of $W1 = 16$, very high completeness ($\gtrsim 80\%$) and relatively low contamination levels ($\lesssim 25\%$) are retained for galaxies and stars. For quasars, the classifier achieved excellent performance of $\sim 95\%$ completeness and $\sim 3\%$ contamination for all the magnitude and extinction ranges.

5 Future prospects

The next step of our project will be to apply the SVM classifier, trained on the WISE×SDSS data, to all-sky data from WISE. This will be presented in the forthcoming paper (Kurcz et al., in prep.). Furthermore, we also plan to extend the present study by examining various SVM kernels, other ML classification methods, as well as the usability of other WISE parameters, for instance through a principal component analysis (cf. Soumagnac et al. 2015). In a longer term, we plan to publicly release thus obtained galaxy, quasar and star catalogues, as we believe that reliable identification of the sources will be of interest for the broader community.

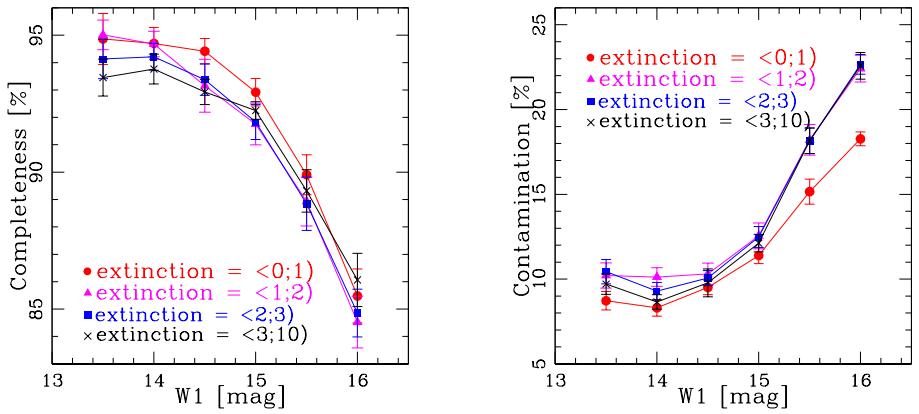

Fig. 5: Dependence of the completeness (left) and contamination (right) on the magnitude $W1$ for all extinction bins for galaxies. Relevant purity levels are 100%− contamination.

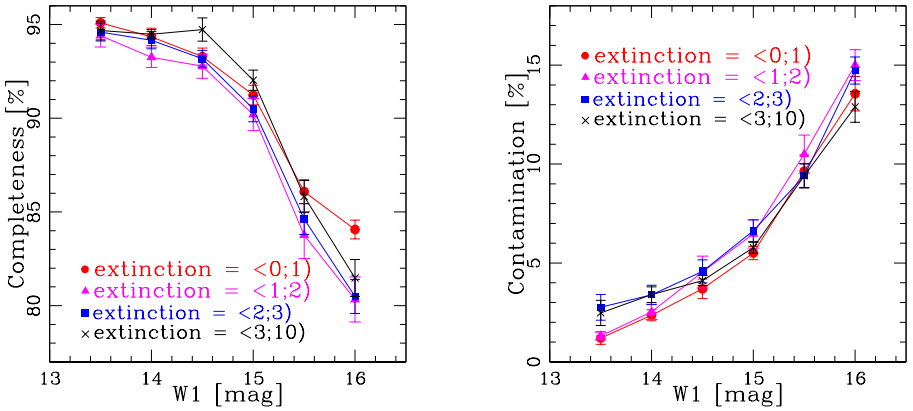

Fig. 6: Dependence of the completeness (left) and contamination (right) on the magnitude $W1$ for all extinction bins for stars.

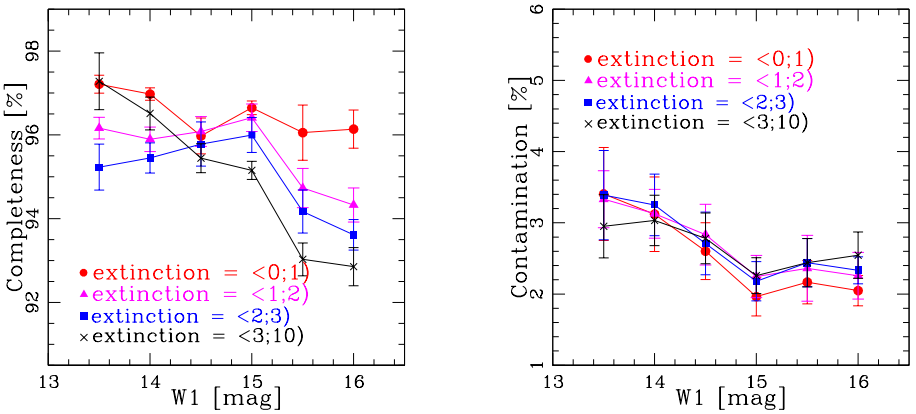

Fig. 7: Dependence of the completeness (left) and contamination (right) on the magnitude $W1$ for all extinction bins for quasars.

Acknowledgements. This work was supported by the Polish National Science Center under contracts no. UMO-2012/07/D/ST9/02785 and UMO-2015/16/S/ST9/00438 (AS). AP was partially supported by the Polish-Swiss Astro Project, co-financed by a grant from Switzerland, through the Swiss Contribution to the enlarged European Union. KM has been supported by the National Science Centre (grant UMO-2013/09/D/ST9/04030).

References

Ahn, C. P., et al., *The Tenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Apache Point Observatory Galactic Evolution Experiment*, ApJS **211**, 17 (2014), 1307.7735

Bilicki, M., et al., *WISE×SuperCOSMOS photometric redshift catalog: 20 million galaxies over 3π steradians*, *ApJS, submitted* (2016)

Cutri, R. M., et al., *Explanatory Supplement to the AllWISE Data Release Products*, Technical report (2013)

Dawson, K. S., et al., *The SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data*, *ArXiv e-prints* (2015), 1508.04473

Ferraro, S., Sherwin, B. D., Spergel, D. N., *WISE measurement of the integrated Sachs-Wolfe effect*, Phys. Rev. D **91**, 8, 083533 (2015), 1401.1193

Jarrett, T. H., et al., *The Spitzer-WISE Survey of the Ecliptic Poles*, ApJ **735**, 112 (2011)

Kovács, A., Szapudi, I., *Star-galaxy separation strategies for WISE-2MASS all-sky infrared galaxy catalogues*, MNRAS **448**, 1305 (2015), 1401.0156

Krakowski, T., et al., *Machine-learning identification of galaxies in the WISE×SuperCOSMOS all-sky catalogue* (in prep.)

Kurcz, A., et al., *Towards an automatic classification of all WISE sources* (in prep.)

Murakami, H., et al., *The Infrared Astronomical Mission AKARI*, PASJ **59**, 369 (2007), 0708.1796

Neugebauer, G., et al., *The Infrared Astronomical Satellite (IRAS) mission*, ApJ **278**, L1 (1984)

Schlegel, D. J., Finkbeiner, D. P., Davis, M., *Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds*, ApJ **500**, 525 (1998), 9710327

Secrest, N. J., et al., *Identification of 1.4 Million Active Galactic Nuclei in the Mid-Infrared using WISE Data*, ApJS **221**, 12 (2015), 1509.07289

Skrutskie, M. F., et al., *The Two Micron All Sky Survey (2MASS)*, AJ **131**, 1163 (2006)

Soumagnac, M. T., et al., *Star/galaxy separation at faint magnitudes: application to a simulated Dark Energy Survey*, MNRAS **450**, 666 (2015), 1306.5236

Stern, D., et al., *Mid-infrared Selection of Active Galactic Nuclei with the Wide-Field Infrared Survey Explorer. I. Characterizing WISE-selected Active Galactic Nuclei in COSMOS*, ApJ **753**, 30 (2012), 1205.0811

Tu, X., Wang, Z.-X., *Classification study of WISE infrared sources: identification of candidate asymptotic giant branch stars*, *Research in Astronomy and Astrophysics* **13**, 323 (2013), 1207.0294

Wright, E. L., et al., *The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance*, AJ **140**, 1868 (2010), 1008.0031