WISE-2MASS all-sky infrared galaxy catalog for large scale structure

András Kovács¹,² and István Szapudi³

¹ Institute of Physics, Eötvös Loránd University, 1117 Pázmány Péter sétány 1/A Budapest, Hungary
² MTA-ELTE EIRSA “Lendület” Astrophysics Research Group, 1117 Pázmány Péter sétány 1/A Budapest, Hungary
³ Institute of Astronomy, University of Hawaii 2680 Woodlawn Drive, Honolulu, HI, 96822, USA

Submitted 2013

ABSTRACT
We combine photometric information of the WISE and 2MASS infrared all-sky surveys to produce a clean galaxy sample for large-scale structure research. Adding 2MASS colors improves star-galaxy separation substantially at the expense of losing a small fraction of the galaxies: 93% of the WISE objects within the $W1 < 15.2$ mag limit have 2MASS observation as well. We use a class of supervised machine learning algorithms, Support Vector Machines (SVM), to classify objects in our large data set. We used SDSS PhotObj table with known star-galaxy separation for a training set on classification, and the GAMA spectroscopic survey for determining the redshift distribution of our sample. Varying the combination of photometric parameters input into our algorithm revealed that $W1_{WISE} - J_{2MASS}$ is a simple and effective star-galaxy separator, capable of producing results comparable to the multi-dimensional SVM classification. The final catalog has an estimated $\sim 2\%$ stellar contamination among 5 million galaxies with $z_{med} \approx 0.17$. This full sky galaxy map with well controlled stellar contamination will be useful for cross correlation studies.

Key words: catalogues – large-scale structure of Universe

1 INTRODUCTION

In recent years, sky surveys have been producing astronomical data with a rapidly accelerating pace resulting in what is commonly called the “data avalanche”. The large quantity of data necessitates automated algorithms for filtering, photometric selection, and estimation observables such as redshifts. Each object in a catalog has multiple properties, thus algorithms have to explore high-dimensional configuration spaces in large, often connected, databases. Such high dimensional spaces can be effectively explored with machine learning techniques, such as the support vector machines (SVM) used in our work.

When analyzing an object catalog, the most fundamental, and often most challenging task is star-galaxy (possibly QSO) separation. A simple separator between stars and galaxies is a morphological measurement, where extended sources are classified as galaxies (Vasconcellos et al. 2011). Morphology, however, looses its power at fainter magnitudes, a problem for wide-field surveys, e.g., Pan-STARRS (Kaiser et al. 2010), Euclid (Amendola et al. 2013), BigBOSS (Schlegel et al. 2011), DES (The Dark Energy Survey Collaboration 2005), and LSST (LSST Science Collaboration et al. 2009). At the fainter end, the most widely used tools for object classification are color-color diagrams: different types of objects will appear in different regions according to the shape of their spectral energy distribution. Classification methods based on color-color selection were employed for star-galaxy separation (Pollo et al. 2010) or for finding special classes of sources, such as high/low redshift QSO, AGN, starburst galaxies, or variable stars (e.g., Richards et al. 2002, Chiu et al. 2005, Stern et al. 2012, Brightman & Nandra 2012, and references therein).

Machine learning techniques, in particular SVMs, are gaining popularity in astronomical data mining and analysis due to their effectiveness and relative simplicity. For instance, Wozniak et al. (2004) analyzed variable sources with an SVM in 5 dimensions constructed from their period, amplitude and three colours; Huertas-Company et al. (2009) estimated morphological properties of near-infrared galaxies using SVM with 12 parameters, while Solarz et al. (2012) created a star-galaxy separation algorithm based on mid and near-infrared colors. Recently, Malek et al. (2013) used VIPERS and VVDS surveys to perform object classification into 3 groups: stars, galaxies, and AGNs. The Photometric Classification Server for the prototype of the Panoramic Survey Telescope and Rapid Response System 1 (Pan-STARRS1) uses SVM as well (Saglia et al. 2013).

While star-galaxy separation can be performed from WISE colors alone (Goto et al. 2012, Kovács et al. 2013) it is at the expense of severe cuts that still are sensitive to contamination to from the Moon and necessitate complex masks. Adding 2MASS observations to WISE cleans up the data significantly. With several open source implementations and computationally modest cost (Fadely et al. 2012), we set out to use the SVM algorithm for separating stars and galaxies in the matched WISE-2MASS photomet-
correlation studies, such as ISW measurements, and galaxy-CMB
ture and cross-correlation studies. At the same time, we will show
stellar streams and Galactic structure in general.
lensing correlations, while the large data sets of stars mayconstrain
ies observed by WISE and 2MASS suitable for large scale struc-
with similar methods (Kovács et al. 2013).
12 and 22
lite, which surveyed the sky at four different wavelengths: 3.4, 4.6,
(2006)). We use photometric measurements of the WISE satel-
t and 2-Micron All-Sky Survey (2MASS, Skrutskie et al.
We combine measurements of two all-sky surveys in the in-
2 DATASETS AND METHODOLOGY
The paper is organized as follows. Datasets and algorithms are
with detailed discussion, comparisons, and interpretation.
2.1 Support Vector Machines
SVM designates a subclass of supervised learning algorithms for
classification in a multidimensional parameter space. These meth-
ods include extensions to nonlinear models of the generic (linear)
algorithm developed by Cortes & Vapnik (1995). SVMs carry out
object classification and/or regression by calculating decision hy-
perplanes between sets of points having different class members-
A central concept of SVM learning is the training set, a special
set of objects that supplies the machine with classified exam-
ships. A central concept of SVM learning is the training set, a spe-
object classification and/or regression by calculating decision hy-
machine as implemented in the python package scipy. We found
an SDSS match for 99.4% for the 46,749 WISE-2MASS objects
using a 3" matching radius. As a further refinement, we applied a
W1 \geq 12.0 magnitude cut to exclude bright objects with potentially
problematic SDSS classification.
As an exploratory test, we downloaded 2MASS XSC data
from the same coverage, finding 1,195 galaxies. The WISE-
2MASS sample contains 5,922 objects classified as a galaxy in
SDSS PhotoObj table. We will show that the fraction of the prop-
erily identified galaxies reaches \approx 75% even with our simplest
algorithms, thus we are able to broaden 2MASS XSC signficantly.
The corresponding redshift distribution of the matched ob-
jects classified as galaxies are shown on Figure 1. These redshifts
are provided by matching with the Galaxy and Mass Assembly
(GAMA, Driver et al. (2011)) dataset, at the full GAMA coverage
of 290 deg^{-2}. We used 3" as a matching radius, and found a pair
for 84% of the WISE-2MASS objects, that is in good agreement
with the findings of Goto et al. (2012) and (Kovács et al. 2013).
The estimated median redshift of the WISE-2MASS sample is
z_{median} \approx 0.17.

Figure 1. Estimated redshift distribution of our WISE-2MASS galaxy sam-
ple, after matching to the GAMA spectroscopic dataset. We found a pair
for 84% of the WISE-2MASS galaxies. This sample has $z_{median} = 0.17,$
slightly deeper than the previous full sky WISE galaxy sample produced
with similar methods (Kovács et al. 2013).

ric data. Our principal goal is to create a clean catalog of galax-
ies observed by WISE and 2MASS suitable for large scale struc-
ture and cross-correlation studies. At the same time, we will show
that our selection algorithms are suitable for producing clean stellar
samples as well. The galaxy maps we create are useful for cross-
correlation studies, such as ISW measurements, and galaxy-CMB
lensing correlations, while the large data sets of stars may constrain
stellar streams and Galactic structure in general.
The paper is organized as follows. Datasets and algorithms are
described in Section 2, while our results are presented in Section 3,
with detailed discussion, comparisons, and interpretation.

2 DATASETS AND METHODOLOGY
We combine measurements of two all-sky surveys in the in-
fraed, Wide-Field Infrared Survey Explorer (WISE, Wright et al.
(2010)) and 2-Micron All-Sky Survey (2MASS, Skrutskie et al.
(2006)). We use photometric measurements of the WISE satel-
-lite, which surveyed the sky at four different wavelengths: 3.4, 4.6,
12 and 22 \mu m (W1-W4 bands). Following Goto et al. (2012) and
Kovács et al. (2013) we select sources to a flux limit of $W1 \leq 15.2$ mag to have a fairly uniform dataset. We add 2MASS J, H
and K magnitudes conveniently available in the WISE catalog. We
keep 93% of the WISE objects with $W1 \leq 15.2$ mag that have
2MASS observations. We note that this choice allows us to produce
a deeper catalog than the 2MASS Extended Source Catalog (XSC,
Jarrett et al. (2000)), as proper identification of fainter 2MASS ob-
jects becomes possible.

To apply machine learning techniques, one needs to identify a
"training set", a set of objects with known classification. We
choose a smaller region of Stripe 82 in the Sloan Digital Sky Survey
(SDSS, Abazajian et al. (2009)), deeper than our catalog and loca-
ted at $327.5 < RA < 338.5$ and $-1.25 < Dec < 1.25.$ We per-
formed the cross-matching with the KD-Tree (Bentley (1975) algo-
rithm as implemented in the python package scipy. We found
an SDSS match for 99.4% for the 46,749 WISE-2MASS objects
using a 3" matching radius. As a further refinement, we applied a
W1 \geq 12.0 magnitude cut to exclude bright objects with potentially
problematic SDSS classification.

Apart from kernel functions, SVM offers a whole set of
parametrization choices. We chose ‘C-classification’ because of its
good performance and only two free parameters. $C$ is the cost func-
tion, i.e. a trade-off parameter that sets the width of the margin
between different classes of objects, and this increases the
chance of mis-classifications (Malek et al. 2013). The second pa-
rameter, $\gamma$, determines the topology of the decision surface. A low
value of $\gamma$ sets a rather rigid, and complicated decision boundary, while a value of $\gamma$ that is too high can give a very smooth decision surface causing mis-classifications. We used a free software environment for SVM in python package scikit-learn.

3 DISCUSSION

3.1 SVM outputs

We use the SVM algorithm implemented in python. First we performed tests to tune both the C and $\gamma$ parameters, and found the lowest classification errors with C=10.0, and $\gamma = 0.1$. Then we proceeded to determine the optimal number of parameters for the optimal classification efficiency experimentally. We used 8,000 objects as a “training set”, and 2,000 objects for control, i.e. testing the efficiency of our algorithms. We evoke the terminology of machine learning, and use "True" (T) and "False" (F) labels to distinguish between objects that are classified correctly, and the ones have false identification.

We also define five measures of SVM performance:

- Star Contamination = $\frac{T_S}{T_S+F_G}$
- Galaxy Contamination = $\frac{T_G+F_S}{T_G+F_S}$
- Star Completeness = $\frac{T_S}{T_S+T_G}$
- Galaxy Completeness = $\frac{T_G}{T_G+F_S}$
- Accuracy = $\frac{T_G+F_S}{T_G+F_S}$

We used the following set of colors/magnitudes as input parameters: W1,W1-W2, W2-W3, W3-W4, J-H, H-K, W1-J, W2-H, and W3-K. Initially, we supplied SVM with all possible pairs of this set, and obtained contamination, completeness, and accuracy. As shown in Figure 2 parameter W1-J is an astoundingly good star-galaxy separator. Either alone or combined with any other parameter W1-J guarantees the lowest star contamination, the highest galaxy completeness, and the highest accuracy. For instance, the star contamination for the combination of W1 and W1-J, or H-K and W1-J is as low as $\sim 3\%$, while the galaxy completeness is $\sim 94\%$.

Figure 2. Measures of SVM performance are presented in the case of pairwise and single usage. Color-coded maps illustrate contamination, completeness, and accuracy for every combinations. All subfigures suggest that W1-J is a dominant potency in star-galaxy separation. We note, however, that SVM failed to produce precise results using W1 alone. Every object was classified as a star with that choice, thus we excluded the W1-only case from the analysis. Combinations of W1 and other parameters, however, are preserved, as they produce valuable results.

Figure 3. Measures of SVM performance are shown as a function of SVM parameters. We observed upgrading trends in contamination, completeness, and accuracy for both stars and galaxies. High completeness values for the star sample can be explained by the fact that the sample is dominated by stars, thus False galaxies cannot affect star completeness significantly.
Next we supplied SVM with more parameters. We started with W1-W2 alone, then added one more parameter in each step. Our findings are summarized in Figure 3. We qualitatively confirmed our former results, namely that the combination of WISE and 2MASS parameters increases the SVM performance. For WISE colors only, galaxy completeness is at the level of \(\sim 40\%\), while with all parameters it reaches \(\sim 95\%\). At the same time, star contamination decreased from \(\sim 15\%\) to \(\sim 3\%\). Finally, similar trends are seen for the accuracy parameter, that incremented by \(\sim 10\%\) by adding 2MASS parameters.

### 3.2 SVM vs. color-color and color-magnitude cuts

The findings of the previous subsection suggest that separation of stars and galaxies can be achieved a linear cut on the W1–W1-J color-magnitude plane. Stellar contamination and galaxy completeness are then comparable to that of the multicolor SVM algorithm, but with a simpler and faster method.

Figure 4 shows the estimated stellar contamination and the ratio of used and lost galaxies in the case of different W1-J cuts. We choose W1-J \(\leq -1.7\) for our purposes, as it guarantees the lowest stellar contamination, while \(\sim 75\%\) of the galaxies can be classified as galaxies.

Visualizing this parameter choice, in Figure 5 we show a W1–W1-J, and WISE color-color diagrams for WISE-2MASS objects from subsample of 10,000 objects. Classes are indicated by SDSS in this comparison. We note that a remarkable separation of stars and galaxies can be seen on the left panel of Figure 5. Objects over-plotted with gray crosses enforce our W1 \(\geq 12.0\) magnitude cut, as a larger subsample of SDSS “galaxies” imbedded in the definite stellar locus of this plot. This fact suggests that these objects might have been mis-classified by SDSS, and their usage is unsafe in a training set. We emphasize, that neither our SVM methods nor the W1–W1-J based simple galaxy selection are not affected, since we removed the brightest objects from our sample.

### 3.3 Further comparisons

Next we compare our galaxy sample to that of Goto et al. (2012), and Kovács et al. (2013). While these works used all four observations of all the four WISE bands, we only need W1 from WISE. As a result, artifacts are significantly reduced, most notably the stripe-shaped over-densities at several locations across the sky shown on the upper panel of Figure 6. As it was pointed out by Kovács et al. (2013), the stripes are associated with the scanning strategy of the WISE survey. Different WISE bands have different sensitivities and sky coverage, therefore affect the uniformity of a full sky sam-
Figure 5. Left: A simple star-galaxy separator which uses only W1, and W1-J color. Separation of stars and galaxies is remarkably strong in this parameter space. Right: Color-color plots of the four WISE bands. We show the special galaxy separator cuts applied by Goto et al. (2012). This result illustrates that star-galaxy separation on traditional color-color planes with linear cuts is challenging, if one wants to use a large fraction of the achievable galaxy sample. Both: Gray over-plotted crosses show objects that were removed by the W1 $\geq$ 12.0 magnitude cut. Interestingly, many SDSS “galaxies” lie far from the galaxy locus we have identified here. We note that our W1-J $\leq$ -1.7 galaxy selection cut is not affected by this SDSS subsample.

Figure 7. WISE-2MASS galaxy density map with our ‘WMAP + |b| < 20’ mask.

W1-J $\leq$ -1.7 gives the lowest contamination according to our tests, we selected galaxies with the following query:

\[
\text{w1mpro between 12.0 and 15.2 and } n_{\text{2mass}} > 0 \text{ and } \text{w1mpro} - j_{\text{m_2mass}} < -1.7 \text{ and } \text{glat not between -10 and 10}
\]

After a few hours of running, we downloaded $\sim$ 5 million WISE-2MASS objects from the IRSA website\footnote{http://irsa.ipac.caltech.edu/}. The dataset contained W1,W2,W3 and W4 for WISE, and J, H and K for 2MASS as photometric parameters, and we also downloaded ‘cc flag’ val-

3.4 All-sky galaxy catalog

W1-J cut appears to be a powerful tool for separating stars and galaxies. This is the fastest and simplest option to create a full-sky galaxy map. An SVM run for the full-sky WISE-2MASS sample would last in $\sim$ 7 days using a dual-core laptop. The simple cut can be realized by a query into the WISE-2MASS database.
We focused on creating a full sky galaxy map based on the joint analysis of WISE and 2MASS photometric datasets. Using 2MASS colors add useful information, while $\sim 93\%$ of the WISE objects with $W1 < 15.2$ mag have 2MASS pairs. We performed star-galaxy separation using a class of wide-spread machine learning tools, Support Vector Machines. WISE-2MASS objects were cross-identified with SDSS objects, and available SDSS PhotoObj classification data were used as training and control sets.

Exhaustive testing of the SVM algorithm with different parameters and inputs revealed that a simple $W1-J<1.7$ photometric color cut produces similarly clean data set as the SVM classification, at the expense of loosing a modest fraction of the galaxies. Thus finally we opted for the simpler method, and produced a clean galaxy sample with $\sim 2\%$ stellar contamination reaching $\sim 75\%$ completeness. On both counts, the resulting full sky survey represents significant improvement over previous samples using WISE colors only for selection. Our galaxy catalog we created with the simple $W1-J<1.7$ color cut contains $N_{gal} \approx 5$ million objects, with an estimated star contamination of $1.8\%$.

Further refinements, such Galactic dust corrections, tests of magnitude limits, and creation of masks are left for future work. We plan to make our future, refined galaxy catalogs public. Our ultimate goal, however, is a full-sky WISE-2MASS-SuperCOSMOS matched galaxy map with photometric redshifts, extending the efforts of [Bilicki et al. 2014]. We plan to perform various cosmological tests with this future full sky 3D galaxy catalog.

ACKNOWLEDGMENTS

We take immense pleasure in thanking the support of NASA grants NNX12AF83G and NNX10AD53G. In addition, AK acknowledges support from Campus Hungary fellowship program, and OTKA through grant no. 101666. We thank István Csabai for useful comments improving the SDSS-WISE2MASS matching properties, and target selection. This publication makes use of data products of the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration, and 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 Space Administration and the National Science Foundation. This work also makes use of data products of the GAMA survey and the Sloan Digital Sky Survey.

REFERENCES

Abazajian K. N., Adelman-McCarthy J. K., Agüeros M. A., Allam S. S., Allende Prieto C., An D., Anderson K. S. J., Anderson S. F., Annis J., Bahcall N. A., et al. 2009, ApJS, 182, 543
Amendola L., Appleby S., Bacon D., Baker T., Baldi M., Bartolo N., Blanchard A., Bonvin C., 2013, Living Reviews in Relativity, 16, 6
Bentley J. L., 1975, Communications of the ACM, 18, 509
Bilicki M., Ade P. A. R., Ade J. R. H., Ahn Y., Allam S. V., Anderson E., et al., 2014, ApJS, 215, 18
Bilicki M., Ade P. A. R., Ade J. R. H., Ahn Y., Allam S. V., Anderson E., et al., 2014, ApJS, 215, 18
Goto, K., Ito, Y., Ishihara, T., et al. 2012, ApJS, 199, 26
Goto, K., Ito, Y., Ishihara, T., et al. 2012, ApJS, 199, 26

Figure 8. Distributions of galaxies on color-color plane is shown for [Goto et al. 2012] cuts and galaxies selected in present work. Significant amount of galaxies is identified properly with the $W1-J < -1.7$ galaxy cut.
