Search for unusual objects in the WISE Survey

Aleksandra Solarz[1], Maciej Bilicki[2·1·3] and Agnieszka Pollo[1·4]

1. National Center for Nuclear Research, A. Sołtana 7, 05–400 Otwock, Poland
2. Leiden Observatory, Leiden University, the Netherlands
3. Janusz Gil Institute of Astronomy, University of Zielona Góra, Szafrana 2, 65–516 Zielona Góra, Poland
4. Astronomical Observatory of the Jagiellonian University, Orla 171, 30–001 Kraków, Poland

Automatic source detection and classification tools based on machine learning (ML) algorithms are growing in popularity due to their efficiency when dealing with large amounts of data simultaneously and their ability to work in multidimensional parameter spaces. In this work, we present a new, automated method of outlier selection, based on support vector machine (SVM) algorithm called one-class SVM (OCSVM), which uses the training data as one class to construct a model of *normality* in order to recognize novel points. We tested the performance of OCSVM algorithm on Wide-field Infrared Survey Explorer (WISE) data trained on the Sloan Digital Sky Survey (SDSS) sources. Among others, we found $\sim 40,000$ sources with abnormal patterns which can be associated with obscured and unobscured active galactic nuclei (AGN) source candidates. We present the preliminary estimation of the clustering properties of these objects. We found that the unobscured AGN candidates are preferentially located in less massive dark matter haloes (DMH) ($M_{\mathrm{DMH}} \sim 10^{12.4}\,\mathrm{M_{\odot}h^{-1}}$) than the obscured candidates ($M_{\mathrm{DMH}} \sim 10^{13.2}\,\mathrm{M_{\odot}h^{-1}}$). This result contradicts the unification theory of AGN sources and indicates that the obscured and unobscured phases of AGN activity take place in different evolutionary paths defined by different environments.

1 Introduction

The increasing amount of data that is collected from large digital sky surveys, now reaching several peta-bytes including hundreds of million of celestial objects and thousands of parameters measured for each of the observed sources, forces astronomy into finding new ways of efficient detection, segregation and classification of the collected information.

An additional role that they will play is allowing the astronomers to search for some rare or even new astrophysical objects which otherwise were missed within the surveys. This aspect can be studied by exploring previously uncharted parts of the parameter spaces, like the classical color-color diagrams, where the distribution of already known sources can point to rare outliers. However, more often than not, new sources can hide their existence by mimicking the appearance of the regular sources. With the new, automated methods offered by machine learning (ML) algorithms, it is now possible to work in high-dimensional parameter spaces not only to efficiently create samples of regular sources in large amounts of data, but also to search for undersampled or even new celestial objects. The presented work is aimed at detecting anomalies within the Wide-field Infrared Survey Explorer (WISE, Wright et al., 2010) data set based on a pure training sample of galaxies, stars and quasars selected

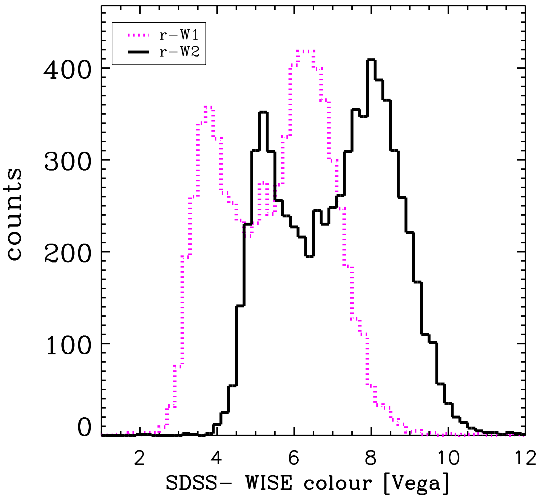

Fig. 1: Optical-infrared color distribution of OCSVM-selected anomalous WISE sources with photometric counterparts in SDSS DR14 database. Solid and dotted lines represent the $r - W2$ and $r - W1$ color distributions. Both colors show similar bimodal behavior indicating existence of at least two source populations within the OCSVM-selected data.

from the cross-match between WISE and Sloan Digital Sky Survey (SDSS, York et al., 2000) catalogs. To find novel sources we use *domain-based novelty detection* method, which is designed to create a boundary based on the structure of the training data set SVM. More commonly, in terms of the usage of SVM for novelty detection, it is known as one-class SVM (OCSVM).

2 The Data

The WISE telescope, launched by NASA in Dec. 2009, scanned the whole sky in four passbands ($W1$—$W4$) covering near- and mid-IR wavelengths centred at 3.4, 4.6, 12 and 23 μm, respectively. Exploration of the publicly available AllWISE catalog (Cutri et al., 2013), which contains over 747 million sources with photometric information, allowed us to test the power of basic artificial intelligence algorithms for anomaly detection in order to obtain information about special objects contained within the dataset. To create the training set, which is the basis of any supervised ML problem, we need to classify manually a representative subset of the data. For this purpose we performed a 1″ radius cross-match between AllWISE dataset with the SDSS DR13 (SDSS Collaboration 2016). This procedure resulted in 2.6 million common sources out of which galaxies comprise 74%, quasars –13%, and stars – 13% of the sample. The second step of the data preparation for a ML procedure is to create a feature vector for each training example, which contains discriminating properties for an object. To that aim we decided to use the $W1$ magnitude measurement, $W1 - W2$ color and a concentration parameter `w1mag13` defined as the difference between flux measurements in two circular apertures for the $W1$ passband in radii equal to 5.5″ and 11.0″ centered on a source (previously used by, e.g., Kurcz et al., 2016).

Table 1: Summary of the obtained correlation function parameters.

	$N_{\rm obj}$	γ	r_0 [Mpc h^{-1}]	b	$M_{\rm DMH}$[M$_\odot$ h^{-1}]
$r - W1 < 5$	743	1.79 ± 0.06	4.57 ± 0.42	1.13 ± 0.10	$10^{12.43}$
$r - W1 > 5$	1212	1.87 ± 0.08	6.96 ± 0.55	1.98 ± 0.13	$10^{13.20}$

3 Method

One of the most popular schemes used for source classification is the Support Vector Machine (SVM, Vapnik 1995). The basic idea behind SVM is that the algorithm is supposed to learn to recognize two (or more) types of objects based on the training examples provided by the supervisor. It uses kernel functions to map the input parameter space into a higher dimensional feature space, where it will search for the best separation hyperplane between the examples of the training points from each category with the biggest margin possible. Then the remaining sources, whose nature is unknown, will have their class assigned based on their relative position to that boundary. It is possible to modify the SVM algorithm as a detection tool, for unrecognizable patterns within the data: instead of using multiple training classes of sources the user has to specify only one class, composed of all the known sources. Then, instead of creating a separation plane, the algorithm will create an enclosed hypershape containing all the known points within the feature space. When the user will apply the remaining unknown sources, all points falling outside of that hypershape will be considered as *anomalies*. This modification is referred to as One-Class SVM (OCSVM) and is perfectly suited for purposes of searching for unusual or unknown sources within large astronomical datasets. For details we refer the reader to Solarz et al. (2017) and references therein.

4 Results

After training the OCSVM algorithm on the AllWISE×SDSS training sample, the full AllWISE data were tested against the created normality model. As a result, we found ∼40,000 sources showing novel properties. The distinguishable property of these sources is their extremely red $W1 - W2$ color (as large as ~ 2 in the Vega system), which means that the sources experience a sharp increase of observed flux with the increase of the observational wavelength. Such behavior and large mid-infrared fluxes can be associated with either warm dust emission or policyclic aromatic hydrocarbon emission lines (characteristic for star-forming galaxies). To confirm the nature of the selected anomalies, we performed a positional cross-match with other publicly available data sets (irrespective of the observational wavelength). We found $\sim 7,000$ counterparts in the photometric part of the SDSS survey, meaning that these sources have optical fluxes measured through the five optical filters, but no spectroscopic redshift information is available. Nevertheless, about $\sim 2,700$ of these sources have their photometric redshifts estimated by Beck et al. (2016). The optical-infrared color distribution shows clearly bimodal behavior indicating that at least two populations of extragalactic sources are contained within this group (see

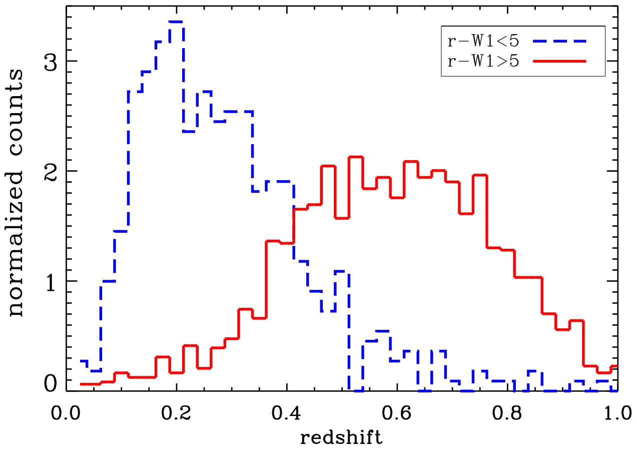

Fig. 2: Photometric redshift distribution for OCSVM-selected AGN candidates divided according to the $r - W1 \sim 5$ criterion into obscured and unobscured sources.

Fig. 1). Similar properties were reported for obscured and unobscured active galactic nuclei (AGN) sources by Donoso et al. (2014).

Two basic types of AGNs, obscured (type-II), and unobscured (type-I), that are being widely observed, are thought to be the result of the orientation of a dust torus around the central black hole. On the other hand, the obscuration of the AGN may rise from larger dust structures like those predicted for major mergers of galaxies (e.g. Hopkins et al. 2006). Simulations by Hopkins et al. (2008) suggested that the dust obscuration could represent a phase of galaxy evolution when a central black hole cannot produce enough accretion luminosity to eject the surrounding material. One of the tests which can provide an answer to this problem is the measurement of the obscured and unobscured AGN clustering as it allows for measurement of the mass of the parent dark matter halo (DMH). If the unification theory is correct, then the two AGN types should appear in similar environments (i.e. similarly massive DMHs). To test this theory we performed a clustering analysis of the two types of AGN sources found through OCSVM analysis; we divided the sample into obscured and unobscured AGNs based on $r - W1 \sim 5$ criterion (cf. Tab. 1). To estimate the angular correlation function we used objects appearing in the northern hemisphere only, as the SDSS coverage is much larger there, and therefore the number of sources available for clustering measurements is greater. The OCSVM-selected AGN candidate sample is not based on any spectroscopic data, only photometric information is available – for that reason the sources used in this work have been never before used for measurements in the large scale structure context.

We used the Landy & Szalay (1993) estimator to evaluate the angular 2-point correlation function and used jack-knife resampling of 32 subsamples to evaluate the errors using full covariance matrix modeling. Usually a correlation function follows the power-law $\omega(\theta) = A_\omega \theta^{1-\gamma}$, where A_ω is the measurement of the correlation strength and γ indicates its scale dependence. Using the measurements of the angular clustering, we can infer the 3-dimensional clustering properties based on the known redshift distribution (shown in Fig. 2) via Limber's equation (Limber, 1954). The obtained results are presented in Fig. 3, for obscured and unobscured

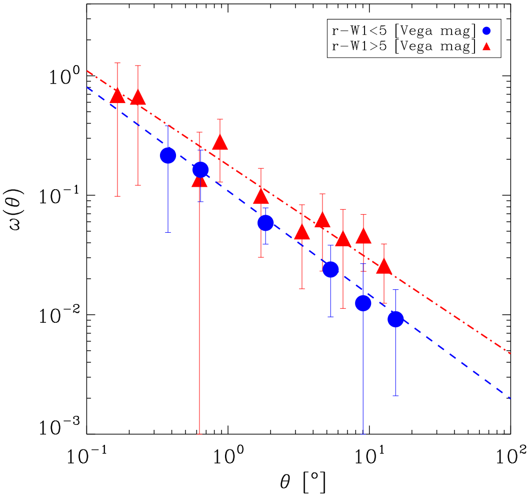

Fig. 3: The angular correlation function for two AGN samples: obscured (marked by red triangles) selected by $r - W1 > 5$ [Vega mag] and unobscured sources (marked by blue circles) selected by $r - W1 < 5$ [Vega mag] cuts. Dashed and dash-dotted lines represent the power-law fit to the correlation function.

AGN candidates. Then, to relate the source clustering to dark matter clustering, it is possible to use a bias parameter – a quantity which describes the differences between the clustering of baryonic field and the underlying mass distribution, i.e. $b^2(r, z, M) = \xi_{\rm g}(r, z, M)/\xi_{\rm m}(r, z)$, where $\xi_{\rm g}(r, z, M)$ is the correlation function of the investigated source population and $\xi_{\rm m}(r, z)$ is the dark matter correlation function. For the details of the calculations we refer the reader to Peebles (1980) and references therein. In Fig. 4, we show the linear bias evolution derived from Sheth & Tormen (1999) formalism for varying minimum DMH mass thresholds.

We found that the OCSVM selected samples of AGN candidates reside in different environments: while the unobscured AGNs at $\langle z_{\rm phot}\rangle \sim 0.26$ are found in haloes which in the present-day Universe reach $\log(M/{\rm M}_{\odot}{\rm h}^{-1}) \sim 12.47$, the obscured sources at $\langle z_{\rm phot}\rangle \sim 0.56$ inhabit haloes of today's $\log(M/{\rm M}_{\odot}{\rm h}^{-1}) \sim 13.20$. The unobscured AGN halo mass is in excellent agreement with the previous works of Donoso et al. (2014) (for obscured and unobscured AGN found in WISE×COSMOS surveys) and Ross et al. (2009) (for SDSS optical quasars), who reported that $\log(M/{\rm M}_{\odot}{\rm h}^{-1}) \sim 12.3$ for type I AGN.

This difference in DMH mass for both AGN types could be a result of the flux-limited nature of the source selection: sources appearing at higher redshifts must be intrinsically brighter to appear within the detection limit of the survey. Objects with higher luminosity are found to have stronger clustering signal than the faint ones (e.g. Zehavi et al., 2011), which could explain the varying DMH masses between the obscured and unobscured AGNs. On the other hand, the merger-driven evolutionary scenario assuming that the AGN obscuration is preceding the unobscured phase of the AGN evolution could explain the fact that the obscured AGN are preferentially found in denser environments than unobscured ones. These findings are contradictory to the AGN unification theory which assumes that the difference between the

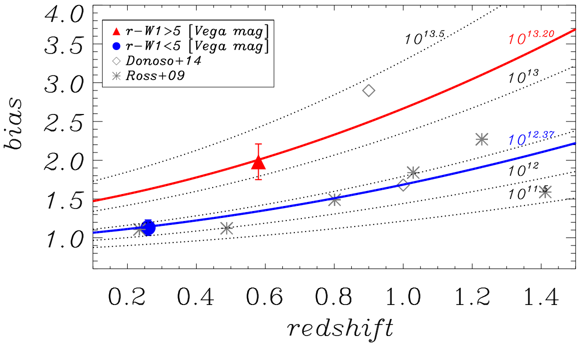

Fig. 4: Linear bias as a function of photometric redshift for obscured (red circle) and unobscured (blue circle) OCSVM-selected AGN candidates. Dashed curves represent the theoretical linear halo bias evolution of DMH of minimal masses from $10^{11.5}$ to $10^{13.5}$ (bottom to top). As a reference we show results from the literature: diamonds from Donoso et al. (2014), asterisks from Ross et al. (2009).

two AGN types is based solely on the orientation of the dusty torus.

Acknowledgements. This publication makes use of the data products from 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. AS has been supported by Polish National Science Centre grant No. UMO-2015/16/S/ST9/00438, MB and AP – by UMO-2012/07/D/ST9/02785.

References

Beck, R., et al., *MNRAS* **460**, 1371 (2016)

Cutri, R. M., Wright, E. L., Conrow, T., Fowler, J. W. e. a., Technical report (2013)

Donoso, E., Yan, L., Stern, D., Assef, R. J., *ApJ* **789**, 44 (2014)

Hopkins, P. F., Hernquist, L., Cox, T. J., Kereš, D., *ApJS* **175**, 356-389 (2008)

Hopkins, P. F., et al., *ApJS* **163**, 1 (2006)

Kurcz, A., et al., *A&A* **592**, A25 (2016)

Landy, S. D., Szalay, A. S., *ApJ* **412**, 64 (1993)

Limber, D. N., *ApJ* **119**, 655 (1954)

Peebles, P. J. E., The large-scale structure of the universe (1980)

Ross, N. P., et al., *ApJ* **697**, 1634 (2009)

SDSS Collaboration (2016), `arXiv:1608.02013`

Sheth, R. K., Tormen, G., *MNRAS* **308**, 119 (1999)

Solarz, A., et al., *A&A* **606**, A39 (2017)

Vapnik, V. N., The nature of statistical learning theory, Springer-Verlag New York, Inc., New York, NY, USA (1995)

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K. e. a., *AJ* **140**, 1868-1881 (2010)

York, D. G., Adelman, J., Anderson, J. E., Jr., SDSS Collaboration, *AJ* **120**, 1579 (2000)

Zehavi, I., et al., *ApJ* **736**, 59 (2011)