Astronomical Data Analysis Software and Systems XXVII
ASP Conference Series, Vol. 522
Pascal Ballester, Jorge Ibsen, Mauricio Solar, and Keith Shortridge, eds.

Novelty Detection in Search of Anomalous Objects within the WISE All Sky Survey

Solarz A.,[1] Bilicki M.,[2,1] and Pollo A.[1,3]

[1]*National Center for Nuclear Research, Warsaw, Poland;*
aleksandra.solarz@ncbj.gov.pl

[2]*Leiden Observatory, Leiden University, the Netherlands*

[3]*The Astronomical Observatory of the Jagiellonian University, Poland*

Abstract. Surveying the previously unexplored areas of the sky guarantees delivering information about unexpected sources whose existence or properties cannot be anticipated. Novelty detection algorithms allow to efficiently search for such objects. In this work we present the ability of one-class support vector machines (OCSVM) algorithm in service of novel source extraction within the infrared AllWISE catalog covering the whole sky.

1. Introduction

Sky Surveys are designed to deliver homogeneous and complete datasets of astronomical sources which are preferably spanning over many cosmic epochs. Later on, the collected data serves as a very foundation for any subsequent scientific analysis, like drawing general conclusions about the bulk of observed objects. Additionally, sky surveys offer a way of searching for rare and unusual sources. For example, studying previously uncharted parts of the parameter spaces (like the color-color diagrams), can reveal the position of rare outliers. However, it is possible that the new sources can mimic the appearance of regular objects. Machine Learning algorithms offer a new, automatic and highly efficient way to exploit a high-dimensional parameter space not only to identify abundant sources in large data-sets but also to find less common or even unexpected astronomical objects. This work is focused on detecting anomalies within the Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010) data set. The training sample of galaxies, stars and quasars is created by performing a positional cross-match between WISE and Sloan Digital Sky Survey (SDSS, York et al. 2000) catalogs. In order to search for novel sources we exploit a *domain-based novelty detection* modification of the classical Support Vector Machine (SVM) algorithm referred to as *one-class SVM* (OCSVM).

2. The Data

The WISE telescope was designed to observe the whole sky in near- and mid-IR wavelengths centered at 3.4 ($W1$), 4.6 ($W2$), 12 ($W3$) and 23 ($W4$) μm. The AllWISE catalog

(Cutri et al. 2013) contains over 747 million sources – due to its large size we can exploit the gathered photometric information about the detected objects to evaluate the performance of artificial intelligence algorithms for anomaly detection tasks. As input, the supervised machine learning problems require a predefined training set, on which they learn to recognize 'usual' data patterns. For that reason it is necessary to manually classify a subset of the data with known characteristics. To this aim we searched for AllWISE source counterparts within 1" radius within SDSS DR13 (SDSS Collaboration 2016) and therefore created a subsample of AllWISE sources with a spectroscopic class confirmation (referred to as AllWISE×SDSS). The created training set is composed of 2.6 million common sources, out of which 74% are galaxies, 13% are quasars and 13% are stars. Next, we need to define a feature vector for each training object, which should preferably include most characteristic properties for a source. For that purpose we use the $W1$ magnitude measurement, $W1 - W2$ color and a concentration parameter `w1mag13` calculated as the difference between flux measurements in two circular apertures in 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).

3. Method

Currently, one of the most popular schemes used for source classification is the Support Vector Machine (SVM, Vapnik 1995). The SVM algorithm is designed to learn how to recognize two (or more) types of objects based on the training examples provided by the supervisor. First, the algorithm maps the input parameter space into a higher dimensional feature space based on kernel functions and then, it searches for the best separation hyperplane between the examples of the training points from each category with the biggest margin possible. Then the unknown sources will be classified based on their relative position to that boundary. One major caveat of the SVM algorithm is that, in the original form, it cannot deal with the data from the general set with patterns unseen during the training process, which results in contaminated output samples of celestial objects. It is possible to modify the SVM algorithm as a tool for detection of patterns within the data that were not included in the training: here the user has to specify only one class, which has to be composed of all the known sources. Then, the algorithm will find contain all the known sources within an enclosed hyper-shape within the feature space. After the algorithm is trained, all new sources which will fall outside of that hyper-shape will be considered as *anomalies*. This modification is referred to as One-Class SVM (OCSVM) and can be efficiently used as a tool for 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, ∼40 000 sources showed novel properties. The distinguishable property of these sources is their extremely red $W1 - W2$ color (~ 2), meaning 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 emis-

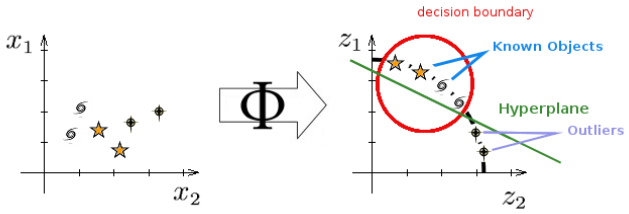

Figure 1. Schematic representation of the OCSVM algorithm operation (Solarz et al. 2017).

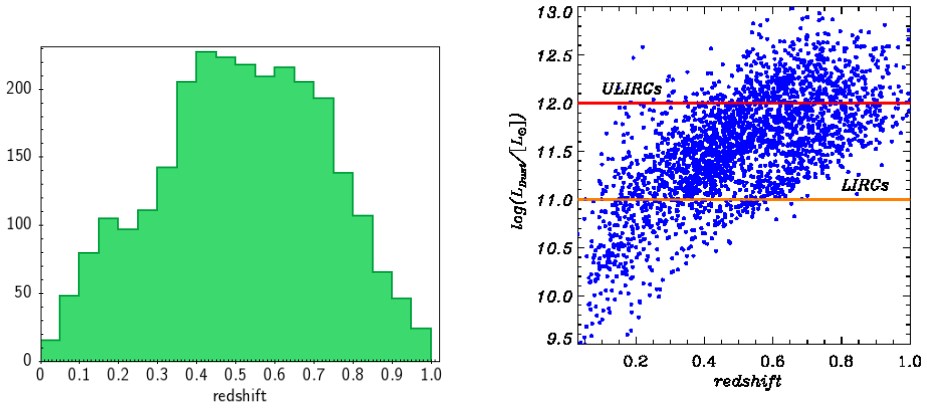

Figure 2. *Left:* The photometric redshift distribution for sources matched with the Beck et al. (2016) catalog. *Right:* L_{TIR} vs redshift for OCSVM anomalous sources.

sion or policyclic aromatic hydrocarbon emission lines (characteristic for star-forming galaxies). To confirm the nature of the selected anomalies we performed a positional crossmatch with other publicly available data sets (irrespective of the observational wavelength). We found $\sim$ 7000 counterparts in the photometric part of the SDSS survey, meaning that these sources have optical fluxes measured through the five optical filters, but no redshift information is available. Nevertheless, about $\sim$ 2700 of these sources have their photometric redshift measured by Beck et al. (2016). With photometric redshifts (shown in left panel of Fig. 2) and optical and infrared photometry it was possible to perform a preliminary spectral energy distribution (SED) fitting to find rough estimations of the physical parameters of these sources. For that purpose we used the CIGALE code (Noll et al. 2009), which incorporates galaxy and AGN templates spanning over the wavelength ranges from ultraviolet to far infrared. Examples of the fits are presented in Fig. 3. We found that $\sim$ 1600 sources can be classified as luminous infrared galaxies (LIRGs), as their total infrared luminosity was estimated to be $L_{TIR} > 10^{11}[L_{\odot}]$. $\sim$ 600 sources showed even larger IR luminosities, $L_{TIR} > 10^{12}[L_{\odot}]$, which places them in the ultraluminous infrared galaxy (ULIRGs) category (shown in right panel of Fig. 2). LIRGs and ULIRGs are thought to experience violent episodes of star formation, due to their large dust content. What is more, according to the Fritz et al. (2006) models, we estimate that the AGN contribution to the IR emission from

the OCSVM anomalies exceeds 50%. The connection between the star-formation and AGN activity is not well understood yet. The samples of the sources exhibiting emission from both components are small, and for that reason the newly found sources in the WISE survey can help to significantly increase the amount of such objects to study their properties. However, spectroscopic follow-up observations are necessary to reveal the true nature of these sources.

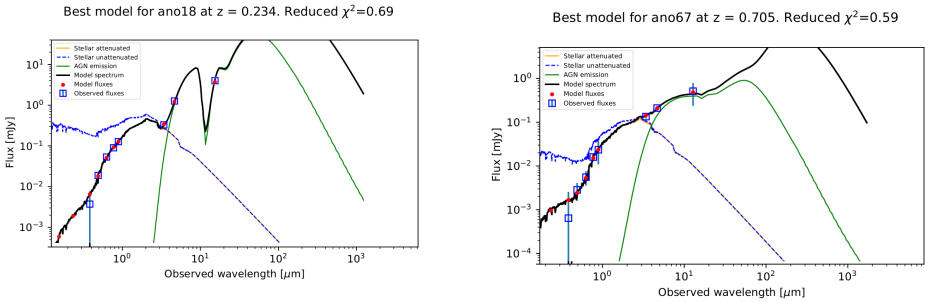

Figure 3. SED fitting examples performed with the CIGALE code for an anomaly classified as LIRG (left) and ULIRG (right) with over 50% AGN contribution to the infrared luminosity.

Acknowledgments. This publication makes use of 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. This research has been supported by National Science Centre grants number UMO-2015/16/S/ST9/00438 and UMO-2012/07/D/ST9/02785.

References

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371. 1603.09708

Cutri, R. M., Wright, E. L., Conrow, T., & Fowler, J. W. e. a. 2013, Explanatory Supplement to the AllWISE Data Release Products, Tech. rep.

Fritz, J., Franceschini, A., & Hatziminaoglou, E. 2006, MNRAS, 366, 767. astro-ph/0511428

Kurcz, A., Bilicki, M., Solarz, A., Krupa, M., Pollo, A., & Małek, K. 2016, A&A, 592, A25. 1604.04229

Noll, S., Burgarella, D., Giovannoli, E., Buat, V., Marcillac, D., & Muñoz-Mateos, J. C. 2009, A&A, 507, 1793. 0909.5439

SDSS Collaboration 2016, ArXiv e-prints. 1608.02013

Solarz, A., Bilicki, M., Gromadzki, M., Pollo, A., Durkalec, A., & Wypych, M. 2017, A&A, 606, A39. 1706.06389

Vapnik, V. N. 1995, The Nature of Statistical Learning Theory (Berlin, Heidelberg: Springer-Verlag)

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868-1881. 1008.0031

York, D. G., Adelman, J., Anderson, J. E., Jr., & SDSS Collaboration 2000, AJ, 120, 1579. astro-ph/0006396