Next Generation Facilities
Proceedings IAU Symposium No. 341, 2018
M. Boquien, E. Lusso, C. Gruppioni & P. Tissera, eds.

doi:10.1017/S1743921319002576

Automatic classification of sources in large astronomical catalogs

Agnieszka Pollo[1,2], Aleksandra Solarz[2], Małgorzata Siudek[3,4,2], Katarzyna Małek[2,5], Maciej Bilicki[6,4], Tomasz Krakowski[2], Tsutomu Takeuchi[7] and the VIPERS team[8]

[1]Astronomical Observatory of the Jagiellonian University, ul. Orla 171, 30-001 Cracow, Poland
email: agnieszka.pollo@gmail.com

[2]National Center for Nuclear Research,ul. A. Sołtana 7, 05-400 Otwock, Poland

[3]IFAE, The Barcelona Institute of Science and Technology, 08193 Bellaterra (Barcelona), Spain

[4]Center for Theoretical Physics, PAS, al. Lotników 32/46, 02-668, Warsaw, Poland

[5]Aix Marseille Univ. CNRS, CNES, LAM Marseille 13388, France

[6]Leiden Observatory, Leiden University, P.O. Box 9513, NL-2300 RA Leiden, The Netherlands

[7]Nagoya University, Furo-Cho, Chikusa-ku, Nagoya 464-8602, Japan

[8]listed at the end of this proceedings

Abstract. In this paper we address two questions related to data analysis in large astronomical datasets, and we demonstrate how they can be answered making use of machine learning techniques. The first question is: how to efficiently find previously unknown or rare objects which can be expected to exist in big data samples? Using the largest existing extragalactic all-sky survey, provided by the WISE satellite, we demonstrate that, surprisingly, *supervised* classification methods can come to aid. The second question is: having a sufficiently large data sample, how can we look for new optimal classification schemes, possibly finding new and previously unknown classes and subclasses of sources? Based on the VIPERS cutting-edge galaxy catalog at redshift $z > 0.5$, we demonstrate that *unsupervised* classification methods can give unexpected but physically well-motivated results.

Keywords. surveys, galaxies: statistics, quasars: general

1. Introduction

Currently, astronomy is dealing with increasingly larger data samples, and even bigger data, like those from the Large Synoptic Survey Telescope (LSST), are coming soon. Such data open new possibilities: both for discoveries of previously unknown rare classes of objects and for understanding the global properties of the known sources, as they provide samples allowing for much more refined statistical treatment. This new situation on the market of astronomical data, coupled with increasing capabilities of modern computers, resulted in a boost of different machine learning and data mining methods applied in the field of astronomy over the last few years.

Very broadly speaking, machine learning based classification methods can be divided into two main families: *supervised* methods, where we know a priori what sources we expect to find and we use some known datasets to train algorithms to look for them, and *unsupervised* methods which look for separate clusters or groups in the data based on similarity of their properties in a given feature space. In this paper, we present two examples of applications of *semi-supervised* (a hybrid between supervised and unsupervised) and

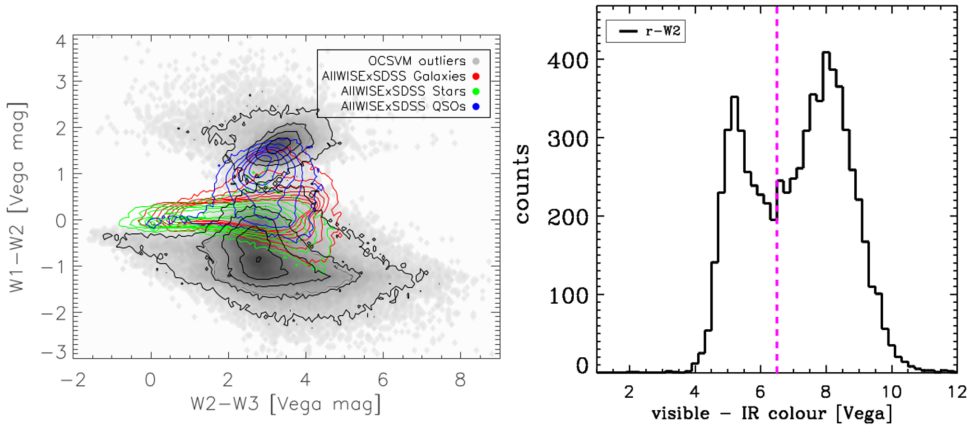

Figure 1. Classification of the AllWISE sources by the OCSVM method. *Left panel*: infrared color-color $W1-W2$ vs $W2-W3$ distribution of AllWISE sources: in addition to objects following the pattern of known galaxies, stars and AGN, a large number of anomalous sources was found. *Right panel*: histogram of the visible-infrared $r-W2$ color of $\sim 7,000$ anomalous objects identified as genuine astrophysical sources with counterparts in the photometric part of the SDSS catalog; the preliminary analysis indicates that they form two groups of dusty (redder) and non-dusty (bluer) AGN.

unsupervised machine learning techniques to search for new unknown classes of sources and new classification schemes for known (and unknown) sources.

2. Supervised methods in search for novel sources

The largest presently existing all-sky extragalactic catalog was created based on the near- to mid-infrared data gathered by the Wide-field Infrared Survey Explorer (WISE, Wright *et al.* 2010) in four passbands ($W1$, $W2$, $W3$, $W4$) centered at 3.4, 4.6, 12 and 23 μm, respectively. It contains over 747 million sources in its AllWISE data release. We can reasonably expect that among such a wealth of data new, rare and previously unclassified categories of objects should be contained. The question is – how to find them?

Solarz *et al.* (2017) applied a *semi-supervised* machine learning method, called one-class support vector machines (OCSVM, Schölkopf *et al.* 2000), in order to separate all *known* classes of objects from those which do not fit to any known pattern in the considered parameter space. A model of expected data was trained based on 2.6 million sources present in the spectroscopic SDSS DR13 database, and found also in AllWISE. As shown in the left panel of Fig. 1, among the AllWISE data with no spectroscopic SDSS counterparts, the OCSVM algorithm found a large number of sources fitting well the pattern corresponding to known stars, galaxies and quasars. However, it also identified two big groups of *anomalous* sources. The largest among these two groups was found to contain mainly artifacts, such as objects with spurious photometry due to blending. It should be stressed that these artifacts were not detected previously by any other cleaning algorithms. Thus, the OCVSM method was proven to be an efficient tool for cleaning large catalogs. However, even more importantly, a second, smaller group of anomalies appeared to contain $\sim 40,000$ real sources of genuine astrophysical interest, among them a sample of heavily reddened Active Galactic Nuclei (AGN) / quasar candidates distributed uniformly over the sky and in a large part absent from other WISE-based AGN catalogs. The $r-W2$ color distribution of $7,000$ among these sources having counterparts in the photometric part of the SDSS survey is presented in the right panel of Fig. 1; two groups

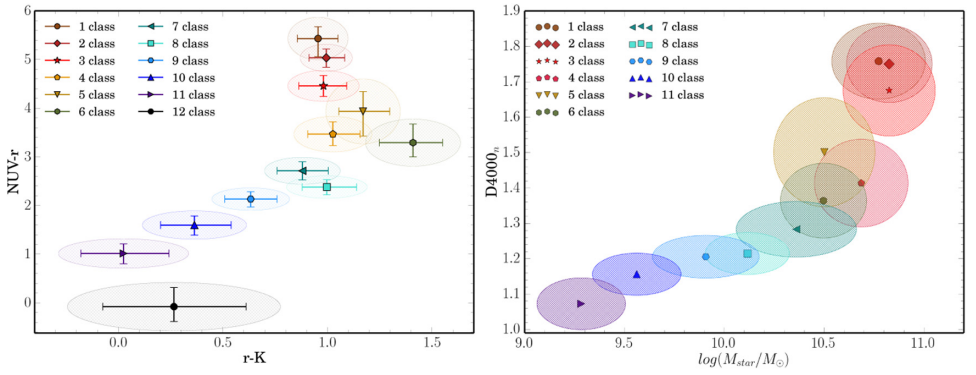

Figure 2. Classification of the VIPERS galaxies by the FEM algorithm. *Left panel*: projection of the FEM classes on the $NUV - r$ vs $r - K$ color-color diagram. The error bars correspond to the first and third quartile of the distribution, while the semi-axes of the ellipses to the median absolute deviation. *Right panel*: median value of the $D4000$ spectral feature as a function of a median stellar mass of the FEM classes. Again, the two semi-axes of the ellipses correspond to the median absolute deviation of the distributions.

of "red" and "blue" sources presumably correspond to obscured and unobscured AGN, respectively. The follow-up observations of these sources are now on-going.

3. Unsupervised methods in search for new classification schemes

The VIMOS Public Extragalactic Redshift Survey (VIPERS; Scodeggio *et al.* 2018) contains ∼90,000 spectroscopically measured galaxies at $z \geqslant 0.5$, covering a large volume of $\sim 5\text{x}10^7 h^{-3}\text{Mpc}^3$ with an effective spectroscopic sampling (∼46%), which makes it a state-of-the-art equivalent of local surveys but at $z \sim 1$.

Siudek *et al.* (2018a) applied a Fisher Expectation-Maximization (FEM) *unsupervised* algorithm (Bouveyron & Brunet 2018) to classify VIPERS galaxies in a parameter space of 12 rest-frame magnitudes and spectroscopic redshift. The FEM algorithm has automatically distinguished 12 classes: 11 classes of galaxies at $0.5 \leqslant z \leqslant 1.2$ and an additional class of outliers consisting mostly of broad-line AGNs. The first broad division was into red (passive), green (intermediate), and blue (star-forming) galaxy populations. A further sub-division yielded three red, three green, and five blue galaxy classes. A joint analysis based on standard statistical criteria (BIC, AIC, ICL), a flow chart of the subsequent divisions performed by the algorithm in the consecutive steps, and a posteriori checks of physical properties of galaxies in resultant FEM classes, allowed for the conclusion that the 11 galaxy classes obtained reflect indeed different galaxy subpopulations, and that the transition of galaxy properties between subsequent classes is not continuous. In other words, the FEM classes can be treated as physically different subcategories of galaxies. The differences between classes are seen not only in galaxy rest-frame magnitudes or colors, but they are also reflected by the properties which were not used for classification neither directly nor indirectly, like spectral lines, shapes or sizes. For example, one among the classes of passive galaxies contains galaxies significantly more compact than two other red classes. As shown in Fig. 2, FEM classes follow the sequence from the earliest to the latest types, which is reflected both by their colors (left panel), and physical as well as spectroscopic properties (right panel).

Can such a fine classification be reproduced based on photometric data only? The question is timely, having in mind huge photometric sky surveys which are expected in the near future, like the LSST. Siudek *et al.* (2018b) extended the previous FEM classification of VIPERS galaxies this time making use of *photometric* redshifts only, and

corresponding rest-frame magnitudes based on these photometric redshifts and apparent magnitudes. Under these conditions, the FEM algorithm effectively distinguished three main red, intermediate, and blue galaxy classes, and the agreement with the corresponding classes based on the spectroscopic data was very high: 92%, 84% and 96%, respectively. Moreover, most of the subclasses were also successfully recovered. The only exception was one of the star-forming classes, containing dusty star-forming galaxies, which in the photometric classification merged with other groups of star-forming galaxies. It was thus proven to be the most sensitive to the accuracy of the recovery of the rest-frame color properties of galaxies.

4. Summary

As we have shown, both supervised and unsupervised machine learning methods give great promise for providing proper understanding of source properties in the future sky surveys. Combining these methods will become a powerful tool for future data analysis. It can allow first for cleaning the data and then for finding novel sources which do not correspond to any known pattern – a semi-supervised algorithm such as OCSVM can add flexibility and reliability to automated source separation procedures. In the next step, unsupervised methods like FEM can be used for efficient and robust classification of already known and, even more importantly, novel samples of sources.

Acknowledgements

This research has been partially supported by the Polish NCN grants 2017/26/A/ST9/00756; 2016/23/N/ST9/02963; 2018/30/M/ST9/00757 and MNiSW grant DIR/WK/2018/12.

References

Bouveyron, C. & Brunet, C. 2012, *Statistics and Computing*, 22, 301
Schölkopf, B., Williamson, R., Smola, A., *et al.* 2000, *Adv. Neural Inf. Process. Syst.*, 582–588
Scodeggio, M., Guzzo, L., Garilli, B., *et al.* 2018, *Astronomy & Astrophysics*, 609, A84
Siudek, M., Małek, K., Pollo, A., *et al.* 2018a, *Astronomy & Astrophysics*, 617, A70
Siudek, M., Małek, K., Pollo, A., *et al.* 2018b, arXiv:1805.09905
Solarz, A., Bilicki, M., Gromadzki, M., *et al.* 2017, *Astronomy & Astrophysics*, 606, A39
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., *et al.* 2010, *Astronomical Journal*, 140, 6, 1868

Discussion

Fumi Egusa: How are different classes different? Are they really discrete groups or part of a continuously/smoothly changing properties?

Agnieszka Pollo: Different classes selected by the FEM algorithm seem to be indeed well separated in the multidimensional feature space. In particular, most galaxies have very clear class assignment with high probability of belonging to a given class and low second best class membership probability; different classes gather galaxies of different physical properties: for instance, one of the red classes gathers galaxies with UV upturn; one of the intermediate classes contains very dusty objects. Thus, the change in properties between different classes is not smooth.

Fumi Egusa: Comparison with $z = 0$ would be interesting.

Agnieszka Pollo: Yes, indeed, and we are working on such a comparison right now.

The VIPERS Team: U. Abbas[5], C. Adami[6], S. Arnouts[6], J. Bel[6,8], M. Bolzonella[9], D. Bottini[18], E. Branchini[10,11,12], A. Cappi[9,13], J. Coupon[14], O. Cucciati[15,9], I. Davidzon[6,9], G. De Lucia[16], S. de la Torre[6], P. Franzetti[18], B. Garilli[18], B. R. Granett[8], L. Guzzo[8,17], O. Ilbert[6], A. Iovino[8], J. Krywult[30] V. Le Brun[6], O. Le Fèvre[6], D. Maccagni[18], F. Marulli[15,19,9], H. J. McCracken[20], M. Polletta[18], L. A. .M. Tasca[6], R. Tojeiro[23], D. Vergani[24,9], A. Zanichelli[25], A. Burden[23], C. Di Porto[9], A. Marchetti[26,8], C. Marinoni[27,12,28], L. Moscardini[15,19,9], J. A. Peacock[29], M. Scodeggio[18], G. Zamorani[9]

[5] INAF - Osservatorio Astronomico di Torino, 10025 Pino Torinese, Italy,

[6] Aix Marseille Universit'e, CNRS, LAM (Laboratoire d'Astrophysique de Marseille) UMR 7326, 13388, Marseille, France;

[7] Canada-France-Hawaii Telescope, 65–1238 Mamalahoa Highway, Kamuela, HI 96743, USA,

[8] INAF - Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, via E. Bianchi 46, 23807 Merate, Italy,

[9] INAF - Osservatorio Astronomico di Bologna, via Ranzani 1, I-40127, Bologna, Italy,

[10] Dipartimento di Matematica e Fisica, Università degli Studi Roma Tre, via della Vasca Navale 84, 00146 Roma, Italy,

[11] INFN, Sezione di Roma Tre, via della Vasca Navale 84, I-00146 Roma, Italy,

[12] INAF - Osservatorio Astronomico di Roma, via Frascati 33, I-00040 Monte Porzio Catone (RM), Italy,

[13] Laboratoire Lagrange, UMR7293, Université de Nice Sophia Antipolis, CNRS, Observatoire de la Côte dAzur, 06300 Nice, France,

[14] Astronomical Observatory of the University of Geneva, ch. d'Ecogia 16, 1290 Versoix, Switzerland,

[15] Dipartimento di Fisica e Astronomia - Alma Mater Studiorum Università di Bologna, viale Berti Pichat 6/2, I-40127 Bologna, Italy,

[16] INAF - Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, 34143 Trieste, Italy,

[17] Dipartimento di Fisica, Università di Milano-Bicocca, P.zza della Scienza 3, I-20126 Milano, Italy,

[30] INAF - Istituto di Astrofisica Spaziale e Fisica Cosmica Milano, via Bassini 15, 20133 Milano, Italy,

[19] INFN, Sezione di Bologna, viale Berti Pichat 6/2, I-40127 Bologna, Italy,

[20] Institute d'Astrophysique de Paris, UMR7095 CNRS, Université Pierre et Marie Curie, 98 bis Boulevard Arago, 75014 Paris, France,

[21] Universitätssternwarte München, Ludwig-Maximillians Universität, Scheinerstr. 1, D-81679 München, Germany,

[22] Max-Planck-Institut für Extraterrestrische Physik, D-84571 Garching b. München, Germany,

[23] Institute of Cosmology and Gravitation, Dennis Sciama Building, University of Portsmouth, Burnaby Road, Portsmouth, PO1 3FX,

[24] INAF - Istituto di Astrofisica Spaziale e Fisica Cosmica Bologna, via Gobetti 101, I-40129 Bologna, Italy,

[25] INAF - Istituto di Radioastronomia, via Gobetti 101, I-40129, Bologna, Italy,

[26] Università degli Studi di Milano, via G. Celoria 16, 20130 Milano, Ital,

[27] Aix Marseille Université, CNRS, CPT, UMR 7332, 13288 Marseille, France,

[28] Université de Toulon, CNRS, CPT, UMR 7332, 83957 La Garde, France,

[29] SUPA, Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK

[30] Institute of Physics, Jan Kochanowski University, ul. Swietokrzyska 15, 25-406 Kielce, Poland