Source classification in large astronomical catalogs

Agnieszka Pollo
National Centre for Nuclear Research
Astronomical Observatory of the Jagiellonian University
Warsaw-Cracow, Poland

with
Ola Solarz, Kasia Malek, Maciek Bilicki, Gosia Siudek, Tomasz Krakowski, Agnieszka Kurcz, Magda Krupa, Tsutomu Takeuchi, AKARI team, VIPERS team, WISE team
We are now living in the epoch of large datasets – both photometric and spectroscopic.

A priori, we usually do not always know what we observe, and the available information is often very limited.
Machine learning for source classification:

**Supervised** → when we know a priori what sources we expect to find and we can use some datasets for training

→ classification (for separate groups) or regression (for smooth transition)

**Unsupervised** → clustering of sources into previously unknown and unexpected classes
In this talk, I will give a couple of examples of a successful application of both these approaches to the source classification in
AKARI (NIR to MIR satellite sky survey)
WISE (NIR to MIR satellite sky survey)
VIPERS (spectroscopic galaxy survey)
Our main – but not only - tool for classification (in this talk...): Support Vector Machines

Basic idea: to move from classifications based on very limited number of parameters (like color-color plots or line-to-line ratio or sth. similar) to the feature space built from a larger number of parameters.

Objects poorly separated in the two parameter space can get well separated in a multi parameter space, and the problem is easier to linearize.
AKARI Extragalactic Deep Survey

AKARI All Sky Survey
9-180um, 6-bands

NEP Deep
0.38 deg², 2-24um, 9-bands

AKARI Deep Field South
12 deg², 50-180um, 4-bands
Source classification of tricky data: AKARI NEP deep field

0.38 sq. deg deep observations in 9 bands 2-24 um
26 sources identified

What these sources are? Stars, galaxies, AGNs?
Not obvious if you do not rely on comparison to optical data (which for the sake of completeness sometimes you do not want to do...)

See a poster by Artem Poliszczuk for a continuation of these works

Solarz et al. 2012, 2015
AKARI NEP: 24-um selected NEP galaxies
N(z) -> at least 3 different galaxy populations at different redshifts

Population centered at $z \sim 0.6$: dusty
Star forming galaxies, Local LIRGs and ULIRGs

Photometric redshifts obtained with the aid of CIGALE, and calibrated with spectro-zs

Population centered at $z \sim 1.2$: sources with 12.7 um PAH feature and/or 12.8 um NeII emission line

Population centered at $z \sim 2.4$: sources dominated by AGNs and/or sources with 8 um PAH line

Solarz et al. 2012, 2015
AKARI: clustering of 24-μm selected NEP galaxies:

All three populations at different redshifts: different but all strongly clustered.

Differently related to underlying dark matter LSS -> evolution environmental dependence of evolution of these three populations.
Source classification of very large data: WISE

- 750 million sources
- four mid-infrared bands: 3.4, 4.6, 12, and 23 μm
- Great training grounds for all types of machine learning analyses
We used the **support vector machines** algorithm trained on SDSS spectroscopic data.

Current **results for W1<16** Vega (1 mag brighter than WISE flux limit) due to limitations of the training set (practically no SDSS galaxies at W1>16).

**45 million galaxy candidates on ~80% of sky**

Inevitable stellar **contamination** at low latitudes - blending due to 6” WISE beam.

*Kurcz et al. 2016*
• We used the WISExSuperCOSMOS sample (Bilicki et al. 2014, 2016): flux-limited cross-matched sample at $|b| > 10^\circ$ with almost 48 million sources.

• **Result:** 15 million galaxy candidates on $\sim$68% of sky

• Is this approach sufficient?

• **See Ola Solarz’s talk tomorrow!**

*Krakowski et al. 2016*
Goal: 100,000 spectra of galaxies at 0.5<z<1.2


Question: based on the observed colors, but having a well defined training sample based on spectroscopic (VIPERS) data, how well can we pre-classify a sample into galaxies, stars and AGNs at 0.5<z<1.2?
VIPERS-trained SVM classifier for AGNs, stars and galaxies at $z>0.5$

- Trained on almost 20,000 VIPERS sources with the best spectroscopic measurement
- Optical (based on 4 apparent magnitudes in $u'$, $g'$, $r'$ $i'$ bands) and NIR+optical classifiers trained
- NIR measurement dramatically increases the classifier’s accuracy
- Classification pattern which is not obvious from color-color plots

Malek, Solarz, Pollo et al. 2013
Question: having such an unprecedented wealth of spectroscopic data, can we classify galaxies better than just traditional blue-red-green valley galaxies?

Method: unsupervised classification based on a feature space of absolute magnitudes + redshifts.

Siudek et al. (hopefully 2017)
Unsupervised classification of galaxies at z>0.5

Unsupervised classification of VIPERS galaxies based on their distribution in a multidimensional absolute magnitude space
12 dimensions: absolute magnitudes + zspec
→ blind separation (no training sample or hints) into 12 classes, which are well separated e.g. in the 3D diagram.

Method: Fisher expectation maximization algorithm (FEM) -

Choosing an optimal clustering model and number of groups based on the Bayesian Information Criterion (BIC).

Siudek et al. (hopefully 2017)
Unsupervised classification of galaxies at $z>0.5$

Multidimensional approach allows to achieve a better separation, while on the standard 2D color-color diagrams these classes overlap, e.g. red passive galaxies (classes 1, 2, and 3) are not distinguishable on UVJ diagram.

Siudek et al. (hopefully 2017)
Unsupervised classification
stacked spectra

$0.4 < z < 0.5$

Siudek et al. (hopefully 2017)
Summary

- In the epoch of large astronomical datasets, we only started to exploit the possibilities of machine learning-based methods.
- The existing datasets provide a very good training ground for future yet larger sky surveys of different type (LSST, Euclid, SKA...)
- ...but each time we need to adopt the method to the problem we want to address and to the properties of the data in question.