Machine learning in astrophysical data

Tomasz Krakowski[1], Katarzyna Małek[1], Maciej Bilicki[1,2], Małgorzata Siudek[1,3] and Agnieszka Pollo[1]

1. National Centre for Nuclear Research, A. Sołtana 7, 05–400 Otwock, Poland
2. Leiden Observatory, Leiden University, P.O. Box 9513 NL-2300 RA Leiden, The Netherlands
3. Center for Theoretical Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02–668 Warszawa, Poland

Astronomical surveys provide an ever-increasing amount of data that requires time consuming analysis. There are many parameters that can be used to distinguish different types of astronomical objects. Therefore, statistical tools are increasingly used for this purpose. For this reason, it is very important to automate this process. One possibility is to use supervised learning algorithms. We present an exemplary application of such an algorithm: supervised learning algorithm based on support vector machines (SVM) applied for classification of the WISE data. Machine learning algorithms can also have other uses, for example, to study clustering of the data. For instance, to study galaxy properties and evolution it would be advisable to categorize them into groups with similar properties. This is usually done in low dimensional parameter space, i.e. based on color-color plots making use of only three or four properties. The disadvantage of such a solution is the possibility of overlooking subtle differences between groups of galaxies. For this reason, we have attempted to use unsupervised learning algorithms to divide data in the multidimensional parameter space; we present preliminary results of such a classification performed on the data from the VIPERS survey.

1 Introduction

The foundations of astrophysical research is the analysis of the observational data. Over the years, with the increasing technical capabilities, the number of sky surveys and the amount of astrophysical data collected from them has been rapidly increasing. Modern surveys contain millions and billions of sources, with many records corresponding to one source. This means that these data are multidimensional. Therefore, their analysis with traditional methods is becoming more and more difficult.

With the growing number of objects, we are getting increasingly uncertain if the classification of objects introduced a years ago is still valid. We may be sure that it is insufficient – new types of objects are being discovered in every new survey. One of the main scientific motivation to create galaxy catalogs is to study the evolution of galaxies. In this case, however, the quality of our work depends on the correct division of objects into individual groups, corresponding to physical differences between them. The traditional approach is to rigidly define the boundaries that determine the types of objects in each group, based, for example on color-magnitude or color-color plots. Sometimes other properties are used, but in most cases the number is limited to three or four. However, due to the number of parameters defining each object, such an approach does not allow to take into account subtle differences between their more finely defined types. This limits the possibilities of analysis of evolution of different classes of galaxies.

Thus, both the increasing amount of astronomical data and the need for new approach to galaxy classification, based on many parameters become a challenge for modern astrophysical research. A solution we propose is in the application of automatic machine learning based tools for the related tasks. A human being cannot easily deal with billions of data records – but modern computers can, with the aid of machine learning algorithms of different types. In this proceeding, we present two such techniques. One is supervised classification performed with the use of the support vector machine (SVM) based algorithm applied for WISExSCOSxSDSS data (Krakowski et al., 2016). Another one is unsupervised FisherEM algorithm applied for classification of VIPERS galaxies at redshift $z \sim 1$ (Siudek et al., 2018).

Machine learning is in fact a set of algorithms that can be used to analyze data sets. The methods can be roughly divided into two basic types: supervised and unsupervised. Supervised methods require that we pre-define a training set. In this way we instruct the algorithm into which classes we want the sources to be classified. Unsupervised methods, like clustering algorithms, do not require any training set but act more or less blindly, trying to find patterns in the data.

The main advantage of those algorithms, comparing to traditional methods (e.g. sophisticated cuts on color-color planes), is the ability to analyze large sets of multidimensional data. The advantages of these algorithms are also visible in the case of classification of objects lying at the border of two sets. Traditional methods are not as flexible and it is difficult to say to which among the neighboring groups the given object belongs – while automated methods allow us to use, e.g. probabilities of belonging to different groups or to create separate catalogs of outliers (Solarz et al., 2017). In such cases, supervised learning algorithms that use all available parameters allow to obtain purer catalogues with finer classification.

Another possibility of using the machine learning algorithms is to look for new and previously unknown divisions and dependencies in data catalogs. Such a task can be accomplished using unsupervised learning algorithms. Their operation is based on the application and optimization of a given statistical model to describe the data.

2 The data

2.1 All sky WISExSCOS catalogue

The two largest photometric surveys in existence today are: an all-sky survey in the infrared provided by the Wide-field Infrared Survey Explorer (WISE) satellite and SuperCOSMOS, built from various different sky survey plate collections. These two datasets have been combined to create a new photometric catalog covering 70% of the sky. The catalog created in this way contains three main types of objects: galaxies, quasars and stars. However, separating these three classes is not an easy task. This division has so far been usually performed using sophisticated cuts on the color-color planes. The disadvantage of this solution is the risk of erroneous classification of objects, e.g. stars as galaxies or galaxies as quasars.

In the work by Krakowski et al. (2016) we decided that the solution to this problem may be the use of the learning algorithm, in this case supervised machine learning SVM method (Małek et al., 2013), which separates objects using a multidimensional plane in a multidimensional feature space. To make this classification

possible, we needed a training sample containing objects whose identity was known. For this purpose, we used the Sloan Digital Sky Survey (Brescia et al., 2015, hereafter: SDSS) which we cross-matched with our catalogue. Below, we list the key properties of all the data used in this study.

- **WISE.** WISE is a photometric survey of the entire sky in bands centered at 3.4 (W1), 4.6 (W2), 12 (W3) and 23 (W4) μm. We used the second full-sky release, ALLWISE dataset, with almost 750 million sources. Even after rejecting the artifacts, Galactic plane and bulge are dominated by stellar blends, therefore we focused on sky area at Galactic latitude $|b| > 10^{\circ}$, which reduced the sample to 460 million sources. In our work, we decided to use flux limit on W1 < 17 mag. This cut left us with 343 million sources. Based on previous studies, (e.g. Kurcz et al., 2016) we estimated that about 100 million sources were galaxies and quasars. Due to the fact that the WISE does not offer reliable aperture photometry we used *w?mpro* magnitudes (where ? stands for the channel number) which are based on point spread functions. As a proxy for morphological properties we used a concentration parameter defined as:

$$\mathrm{W1mag13} = \mathrm{w1mag\_1} - \mathrm{w1mag\_3}, \tag{1}$$

  where w1mag_1 and w1mag_3 are magnitudes measured in fixed circular apertures of radii $5.5''$ and $11''$, respectively. The w1mag13 parameter has different distributions for point and resolved sources, which was verified against SDSS spectroscopic data. Due to very low sensitivity of band W4 and the fact that W3 band does not affect out classifier, we decided to use only W1 and W2 bands to create the classification parameter space.

- **SuperCOSMOS.** SuperCOSMOS (Hambly et al., 2001, hereafter: SCOS) is a digitized sky survey in three bands B, R, I obtained through automatic scanning of plates. In Krakowski et al. (2016), we used resolved sources, and residual quasars and stars were further removed from the sample. Preselection process followed those by Bilicki et al. (2016). To preserve photometric reliability of the sample we applied two flux limits: $B < 21$ mag, and $R < 19.5$ mag.

- **Cross-matched WISExSCOS photometric sample.** Catalogs WISE and SCOS were paired with a matching radius of $2''$. Resulting cross-matched sample contains almost 48 million sources. All the magnitudes were corrected for extinction (as in Bilicki et al., 2016), to avoid biases in the final catalogue. Because the of lack of training data we did not use areas with very high extinction $\mathrm{E}(B-V) > 0.25$. By applying this cut we removed another 7.2 million sources.

- **Training sample: SDSS DR12 spectroscopic data.** The SDSS is a multi-filter imaging and spectroscopic survey, and its DR12 (data release) includes star, galaxy, and quasar surveys (Bolton et al., 2012). Pairing these sources with our WISExSCOS flux-limited catalogue within $1''$ matching radius gave us over 1 million common objects, among them 95% of galaxies, 2% – stars, and 3% – quasars.

2.2 VIPERS

VIMOS Public Extragalactic Survey (VIPERS, Scodeggio et al., 2018; Pollo et al., 2017) is a spectroscopic survey of galaxies at $z \sim 1$. Its final catalogue consists of 86,775 galaxies with spectroscopically measured redshifts and a wealth of auxiliary information. In Siudek et al. (2018) we used the unsupervised algorithm to introduce a new, more precise than ever before, classification of these galaxies. In this work we used only galaxies with the highest redshift reliability ($\geq$99%), which leaves us with 52,114 galaxies in the redshift range $0.4 < z < 1.3$.

3 Methods and results

Supervised machine learning is a technique in which a classifier must be trained. Therefore a well defined training sample with all types of objects expected in catalog is needed. Then, with the training sample, we can start to train a classifier, which will be subsequently used to classify the main part of the data. During the learning process we need to check the quality of the obtained classifier. This is usually done with the test sample, either separated from the training sample, or defined independently. When the classifier is positively verified, we can classify the main catalog. Below, we summarize the results of the WISExSCOS classification by the SVM-based classifier performed by Krakowski et al. (2016).

3.1 Supervised learning – SVM-based classification of WISExSCOS

In WISExSCOS catalogue objects: galaxies, stars and quasars on the color-color plots are overlapping each other. Therefore, finding simple cut to separate classes of objects is non trivial. The need for a more refined tool is obvious.

For the classification of the WISExSCOS sample we used SVM algorithm, which is a supervised learning method. This algorithm is able to find classification planes between sets of different objects by finding decision boundary. The SVM algorithm maximizes a margin between the closest points of different classes (support vectors). In our case, the SVM was used as a nonlinear classifier. To create a classification multidimensional space we used photometric information: magnitudes, colours and differential aperture magnitude defined by Eq. 1. SVM algorithm transforms input data with a kernel into a higher dimensional feature space, in which separation between different classes is less complex. The classifier is trained using a subset of input data for which identifications are known – as explained above, in our case they were derived from the SDSS DR12 spectroscopic catalog.

The SVM classifier was trained and then run on the full WISExSCOS all sky catalog. As a result, each object was classified into one of three groups: stars, galaxies or quasars, together with probabilities of each object belonging to a particular group. The latter gives the possibility of adapting the data sample to the needs of research carried out on it - some may need a catalog which is as pure as possible, for other application completeness is more crucial. Finally, we obtained a catalog consisting of 16.8 millions of galaxies classified by the SVM.

3.2 Unsupervised learning – FEM-based classification of VIPERS

Unsupervised algorithms classify the input data sample with no a priori assumptions about their identity. However, one can decide into how many groups the sample should be divided, what statistical properties of the groups are required, or – which is usually the case – what is the requested balance between the number of groups and their statistical properties. One of the basic advantages of such an algorithm is the possibility of using multiple dimensions simultaneously to create groups from the input sample of objects. Thanks to multidimensional approach, we can obtain a division of data which was never attempted, or discover previously unknown relationships between input data.

In the case of galaxies at $z \sim 1$, many approaches to classification were tried, usually based on different variations of the color-color plots. One of the best known is the NUVrK diagram (Arnouts et al., 2013), which indeed allows for finer galaxy classification than other color combinations but the results are not yet fully satisfactory. In Siudek et al. (2018), we use FisherEM (Bouveyron & Brunet, 2011, hereafter: FEM) algorithm for classification of VIPERS galaxies. The main task of the FEM algorithm is to find optimal parameters of the best statistical model describing the data. Here we chose a multivariate Gaussian function. The FEM algorithm makes use of so called latent subspace, in which input data are linearly transformed.

For the classification we used 12 rest-frame magnitudes: FUV, NUV, u, g, r, i, z, B, V, J, H, and Ks (obtained from the SED fitting) and spectroscopic redshift. Optimal model and number of classes were selected based on a combination of typical statistical criteria: Bayesian information criterion (BIC, Schwarz, 1978), Akaike information criterion (AIC, Akaike, 1974) and integrated completed likelihood (ICL, Baudry, 2012).

VIPERS data were separated into twelve well-defined groups, which can be sorted into three *usual* categories: (1) passive, (2) intermediate, (3) star-forming. Their further division opens, however, new possibilities for studies of galaxy evolution. Similarly to SVM, the FEM also assigns each object a probability of belonging to a particular class, which can be used to adjust classification for specific application.

Groups from this newly created classification have been thoroughly tested and it was shown that they differ in physical properties and that they are statistically separated. Thus, the obtained division is not artificial. In the received classification we see clearly well-defined sequence of classes ranging from star-forming to passive galaxies. This method demonstrates the possibility of using FEM algorithms to obtain new, more refined classifications of galaxies, which will allow for more accurate tracking of their evolutionary paths.

4 Summary

As shown above, machine learning methods can be successfully used for different types of data samples. The advantage of this approach is a possibility to use all the available parameters, and consequently, a more accurate separation in a multiparameter space. The result of the classification is not only the assignment of individual objects to groups, but also their probability of belonging to all possible groups. It allows to adapt the final catalogs to a specific applications; e.g. at the

expense of the lower completeness of the catalog, we obtain a better purity.

The choice of the method depends on the sample we are dealing with, and the goal we want to reach. Having a large sample of poorly understood data with a very uncertain classification, a choice of supervised method is only natural. We should remember, however, that the final classification will be as good as the training samples are. On the other hand, as shown by Solarz et al. (2017), SVM allows also for a successful selection of outliers - sources with unusual properties. On the other hand, having a sample of already well known properties, we can use unsupervised methods to look for previously unknown classes of objects or correlations between source parameters never found before. Both ways, the analysis of future, much larger datasets, like those provided by LSST, will have to rely on machine learning tools.

Acknowledgements. This work was supported by the Polish National Science Centre through grants UMO-2013/09/D/ST9/04030 (KM, TK, MS), UMO-2012/07/D/ST9/02785 (MB, KM, AP), and UMO-2016/23/N/ST9/02963 (MS).

References

Akaike, H., *IEEE Transactions on Automatic Control* **19**, 716 (1974)

Arnouts, S., et al., *A&A* **558**, A67 (2013)

Baudry, J.-P. (2012), `arXiv: 1205.4123`

Bilicki, M., et al., in R. van de Weygaert, S. Shandarin, E. Saar, J. Einasto (eds.) The Zeldovich Universe: Genesis and Growth of the Cosmic Web, *IAU Symposium*, volume 308, 143–148 (2016)

Bolton, A. S., et al., *AJ* **144**, 144 (2012)

Bouveyron, C., Brunet, C. (2011), `arXiv: 1101.2374`

Brescia, M., Cavuoti, S., Longo, G., *MNRAS* **450**, 3893–3903 (2015)

Hambly, N. C., Davenhall, A. C., Irwin, M. J., MacGillivray, H. T., *MNRAS* **326**, 1315–1327 (2001)

Krakowski, T., et al., *A&A* **596**, A39 (2016)

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

Małek, K., et al., *A&A* **557**, A16 (2013)

Pollo, A., et al., in this volume (2017)

Schwarz, G., *The Annals of Statistics* **6**, 2, 461 (1978), URL `http://www.jstor.org/stable/2958889`

Scodeggio, M., et al., *A&A* **609**, A84 (2018)

Siudek, M., et al., in preparation (2018)

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