Education and Heritage in the era of Big Data in Astronomy
Proceedings IAU Symposium No. 367, 2020
R. M. Ros, B. Garcia, S. R. Gullberg, J. Moldon & P. Rojo, eds.
doi:10.1017/S1743921321000132

Machine learning for the extragalactic astronomy educational manual

Maksym Vasylenko and Daria Dobrycheva

Main Astronomical Observatory of the National Academy of Sciences of Ukraine
27 Akademik Zabolotnyi St., Kyiv, 03143 Ukraine
emails: daria@mao.kiev.ua, vasmax@mao.kiev.ua

Abstract. We evaluated a new approach to the automated morphological classification of large galaxy samples based on the supervised machine learning techniques (Naive Bayes, Random Forest, Support Vector Machine, Logistic Regression, and k-Nearest Neighbours) and Deep Learning using the Python programming language. A representative sample of $\sim 315\,000$ SDSS DR9 galaxies at $z < 0.1$ and stellar magnitudes $r < 17.7^m$ was considered as a target sample of galaxies with indeterminate morphological types. Classical machine learning methods were used to binary morphologically classification of galaxies into early and late types (96.4% with Support Vector Machine). Deep machine learning methods were used to classify images of galaxies into five visual types (completely rounded, rounded in-between, smooth cigar-shaped, edge-on, and spiral) with the Xception architecture (94% accuracy for four classes and 88% for cigar-like galaxies). These results created a basis for educational manual on the processing of large data sets in the Python programming language, which is intended for students of the Ukrainian universities.

Keywords. machine learning, morphological galaxy classification, education

1. Introduction

Morphological classification of galaxies is one of the basic aim for extragalactic astrophysics and observational cosmology because the morphology of galaxies plays a vital role in reflecting the evolutionary history and large-scale structure growing in the Universe. The visual labeling is the most precise method of galaxy morphological classification, which is successfully used to within 2 T-type units of the Vaucouleurs scale before era of big data surveys. Modern galaxy sky survey DES generated 50 TB of data over its six observation seasons (average 2 TB of data each night); SDSS detected 116 TB at all; the forthcoming LSST will detect 30 TB per night. It means that these hundreds of millions of galaxies could be impossible to classify manually, and the human mind is not able to comprehend complex correlations in the diverse space of parameters.

So, the multidimensional analysis is the best tool for determining the various features between different types of galaxies. All that exaggerates the interest to use the alternatives in the form of machine learning (ML), including deep learning (DL), for the automated classification of galaxies by their features (Ball & Brunner (2010), Conselice et al. 2014, Vavilova et al. (2020b,c), Elyiv et al. (2020)). Students as the future professional astronomers should be familiar with modern methods of big data processing. In this sense, the educational manual teaching the ML and DL methods for galaxy morphological classification is a good introduction into the modern course of extragalactic astronomy. Our textbook was prepared taking into account the experience with another manual on the multi-wavelength properties of galaxies and galaxy clusters, first of all,

https://doi.org/10.1017/S1743921321000132 Published online by Cambridge University Press

in optical and X-ray spectral ranges (Chesnok et al. (2009), Vol'Vach et al. (2011), Babyk & Vavilova (2014), Pulatova et al. (2015), Vasylenko et al. (2020)).

2. Experimental results for the educational manual

We have successfully applied a supervised machine learning methods to the automated morphological classification of large galaxy samples. Unlike the most other authors, we paid attention to the visual cleaning of the dataset (Dobrycheva (2013)).The studied sample contains of $\sim 315\,000$ SDSS DR9 galaxies at $z < 0.1$ and stellar magnitudes $r < 17.7^m$ with unknown morphological types (Dobrycheva et al. (2018)).

We apply classical ML methods (Naive Bayes, Random Forest, Support Vector Machine, Logistic Regression, and k-Nearest Neighbours) to binary morphologically classification of galaxies into early E (from elliptical to lenticular) and late L (from spiral $S0a$ to irregular Im/BCG) types. The training sample consisted of 6 163 galaxies, which were randomly selected at different redshifts and with different luminosity. For training the classifier, we used the absolute magnitudes, color indices, and inverse concentration index (Melnyk et al. (2012)). As a result, the Support Vector Machine provides the highest accuracy – 96.4% correctly classified (96.1% early E and 96.9% late L) types. It allowed us to create the Catalogue of morphological types of $\sim 315\,000$ SDSS-galaxies at $z < 0.1$ applying the Support Vector Machine: $\sim 141\,000$ E-type and $\sim 174\,000$ L-type galaxies among them (Vavilova et al. (2017), Vasylenko et al. (2019), Khramtsov et al. (2019)).

Deep learning methods were applied to classify images of SDSS-galaxies into five visual types (completely rounded, rounded in-between, smooth cigar-shaped, edge-on, and spiral) We have retrieved target sample RGB images, composed of gri bands colour scaling, each of $100\times100\times3$ pixels. The convolutional neural network classifier (Vasylenko (2020)), namely Xception, was trained on the images of galaxies from the target sample, matched in the Galaxy Zoo 2 dataset. We provided the data augmentation that was randomly applied to the images during learning. Our method shows the state-of-art performance of morphological classification, attaining $> 94\%$ of accuracy for all classes, except cigar-shaped galaxies $\sim 88\%$ (Khramtsov et al. (2019), Khramtsov et al. (2020)).

It is very important to share the accumulated knowledge with the next generation. Educational manual based on these our results in the Python programming language is intended for students of the Ukrainian universities.

References

Babyk, I., & Vavilova, I. 2014, *Ap&SS*, 349, 415
Ball, N.M., & Brunner, R.J. 2010, *Int. J. Modern Phys. D*, 19, 1049
Chesnok, N.G., Sergeev, S.G., Vavilova, I.B. 2009, *Kinemat. Phys. Cel. Bodies*, 25, 107
Conselice, C.J., Bluck, A.F.L., Mortlock, A. et al. 2014, *MNRAS*, 444, 1125
Dobrycheva, D.V. 2013, *Odessa Astron. Publ.*, 26, 187
Dobrycheva, D.V., Vavilova, I.B., Melnyk, O.V. et al. 2018, *Kinemat. Phys. Cel. Bodies*, 34, 290
Elyiv, A.A., Melnyk, O.V., Vavilova, I.B. et al. 2020, *AA*, 635, A124
Khramtsov, V., Dobrycheva, D.V., Vasylenko, M.Y. et al. 2019, *Odessa Astron. Publ.*, 32, 21
Khramtsov, V., Dobrycheva, D.V., Vasylenko, M.Yu. et al. 2020, *AA* (submitted for review)
Melnyk, O.V., Dobrycheva, D.V., & Vavilova, I.B. 2012, *Astrophysics*, 55, 293
Pulatova, N.G., Vavilova, I.B., Sawangwit, U. et al. 2015, *MNRAS*, 447, 2209
Vasylenko, M. Y., Dobrycheva, D. V., Vavilova, I. B. et al. 2019, *Odessa Astron. Publ.*, 32, 46
Vasylenko, M., Dobrycheva, D., Khramtsov, V., et al. 2020, *Communications of BAO*, 67, 354
Vasylenko, A.A., Vavilova, I.B., & Pulatova, N.G. 2020, *Astron. Nachr.*, 341, 801
Vavilova, I.B., Dobrycheva, D.V., Vasylenko, M.Y. et al. 2020, *arXiv eprints* 1712.08955v2

https://doi.org/10.1017/S1743921321000132 Published online by Cambridge University Press

Vavilova, I., Dobrycheva, D., Vasylenko, M. et al. 2020, in: P. Skoda & F. Adam, *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (Elsevier), p. 307

Vavilova, I., Pakuliak, L., Babyk, I. et al. 2020, in: P. Skoda & F. Adam, *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (Elsevier), p. 57

Vol'Vach, A.E., Vol'Vach, L.N., Kut'kin, A.M., et al. 2011, *ARep*, 55, 608

https://doi.org/10.1017/S1743921321000132 Published online by Cambridge University Press