DOI:http://dx.doi.org/10.18524/1810-4215.2019.32.182538

VERIFICATION OF MACHINE LEARNING METHODS FOR BINARY MORPHOLOGICAL CLASSIFICATION OF GALAXIES FROM SDSS

M. Yu. Vasylenko[1,2], D. V. Dobrycheva[1,3], I. B. Vavilova[1], O. V. Melnyk[1], A. A. Elyiv[1]

[1] Main Astronomical Observatory of the National Academy of Sciences of Ukraine, 27 Akademika Zabolotnoho Str., 03143, Kyiv, Ukraine, *vasmax@mao.kiev.ua*

[2] Institute of Physics of the National Academy of Sciences of Ukraine, 46, Nauka avenu, 03028, Kyiv, Ukraine

[3] Bogolyubov Institute for Theoretical Physics of the National Academy of Sciences of Ukraine,b 14-b Metrolohichna Str., 03143, Kiyv, Ukraine

ABSTRACT. We present a study on the verification of Machine Learning methods to be applied for binary morphological classification of galaxies. With this aim we used the sample of 60 561 galaxies from the SDSS DR9 survey with a redshift of $0.02 < z < 0.06$ and absolute magnitudes of $-24^m < M_r < -19.4^m$. We applied the following classification methods using own code in Python to predict correctly the morphology of Late and Early galaxies: Naive Bayes, Random Forest, Support Vector Machines, Logistic Regression, and k-Nearest Neighbor algorithm. To study the classifier, we used absolute magnitudes M_u, M_g, M_r, M_i, M_z, color indices $M_u - M_r, M_g - M_i, M_u - M_g, M_r - M_z$, and inverse concentration index to the center $R50/R90$.

We compared these new results with previous one made with the KNIME Analytics Platform 3.5.3. It turned out that Random Forest and Support Vector Machine Classifiers provide a highest accuracy, as in the previous study, but with help our code in Python we increased an accuracy from 92.9 % of correctly classified (96% – E and 84% – L) to 94,6% (96,9% – E and 89,7 % – L). The accuracy of the remaining methods also grew by 88% to 93%. So, using these classifiers and the data on color indices, absolute magnitudes, inverse concentration index of galaxies with visual morphological types, we were able to classify 60 561 galaxies from the SDSS DR9 with unknown morphological types and found 22 301 E and 38 260 L types among them.

Key words: galaxies, morphological classification, machine learning.

АНОТАЦІЯ. Подано дослідження щодо верифікації методів машинного навчання, що застосовані для автоматичної бінарної морфологічної класифікації галактик. З цією метою ми використали вибірку 60 561 галактик з цифрового огляду SDSS DR9 із червоними зміщеннями $0,02 < z < 0,06$ та абсолютними зоряними величинами $-24^m < M_r < -19,4^m$. Ми застосували такі методи машинного навчання, використовуючи власний код написаний на Python, щоб правильно визначити морфологію ранніх і пізніх типів галактик: наївний Байєс, випадковий ліс, метод опорних векторів, логістичну регресію, k-найближч их сусідів. Для тренування класифікатора ми використовували абсолютні зоряні величини M_u, M_g, M_r, M_i, M_z, показники кольору $M_u - M_r, M_g - M_i, M_u - M_g, M_r - M_z$ та зворотній індекс концентрації кольору до центру $R50/R90$.

Ми порівняли ці результати з нашими попередніми, які були зроблені за допомогою програмного забезпечення KNIME Analytics 3.5.3. Виявилося, що метод Random Forest і Support Vector Machine також забезпечують найбільшу точність, але за допомогою нашого коду на Python ми підвищили точність з 92,9 % правильно класифікованих (96 % - E і 84 % - L) до 94,6 % (96,9 % - E і 89,7 % - L). Точність решти методів також зросла на 88 % до 93 %. Отже, тренуючи ці класифікатори на даних про показники кольору, абсолютні зоряні величини, зворотній індекс концентрації кольору до центру галактик, ми змогли визначити морфологічні типи вибірки 60 561 галактик з цифрового огляду неба SDSS DR9 і отримали 22 301 E ранніх і 38 260 L пізніх типів галактик.

Ключові слова: морфологічна класифіквція галактик, машинне навчання

1. Introduction

As a result of the fast development of new technologies for the ground-based and space-born telescopes the volume of digital data about space objects (including the extragalactic ones) has grown rapidly in recent decades. The massive volume of data and more and more increasing computing power facilities change the way in how science and technology are managed. This opens up and get challenges into further research in each field, hence, instigating the search for new approaches to process this huge astroinformation resource (see, for example, [Zaane (1999), Srivastava et al. (2012), Ivezic et al. (2014), Al-Jarrah et al. (2015), Vavilova (2016)]).

Due to the new astronomical observational surveys, their data collection is available online in the form of big science databases in all ranges of the electromagnetic spectrum: Fermi-GLAST [Acero et al. (2015)] in gamma, ROSAT [Voges et al. (2000)] and XMM-Newton [Rosen et al. (2016), Pierre et al. (2016)] in X-ray, GALEX [Lee et al. (2011)] in ultraviolet, WISE [Wright et al. (2010)] and 2MASS [Skrutskie et al. (2006)] in infrared, Extragalactic Radio Continuum Surveys [Norris et al. (2017)] or Discrete Radio Source Surveys [Braude et al. (2002)], zCOSMOS – deep sky survey [Scoville et al. (2007)], deep surveys with the Hubble Space Telescope, and SDSS – Sloan Digital Sky Survey [Gunn et al. (1998)] in optical ranges as well as other surveys.

Since 2000, the Sloan Digital Sky Survey (SDSS) collected the more data that had been amassed in the entire history of astronomy [Blanton et al. (2017)]. Now, its archive contains of about 170 terabytes of information. In this context, the astronomers, who are directly involved in the SDSS, identified the problem of the automated morphological galaxy classification as one of the extremely actual task. Machine learning methods (MLM) are able to uncover hidden relations between observed data (e.g., galaxy parameters and images) and physical properties of galaxies. First of all, we mention several works related to the morphological classification of galaxies from the SDSS such as [Andrae et al. (2010), Dobrycheva et al. (2017), Dobrycheva et al. (2018), Dominguez et al. (2018), Barchi et al. (2019)] as well as to the visual classification, ZOO project, such as [Banerji et al. (2010)], or for radio galaxies with AGNs [Zhixian et al. (2018)].

For the first time we have introduced and applied the high-order 3D Voronoi tessellation method for the identification of low-populated galaxy systems from a volume-limited SDSS DR5 to estimate environment effects [Vavilova et al. (2005), Elyiv et al. (2009)] and binary morphological content [Vavilova et al. (2009)] of 6 786 galaxies with $3000km/s < V_{LG} < 9500km/s$ containing in these systems.

After enlarging the sample to 317 018 galaxies with these radial velocities from the SDSS DR9, we applied multi-parametric diagram and visual inspection to get the automated galaxy morphological classification. Namely, as for the photometry parameters diagrams we used a well-known fact that galaxy morphological type is correlated with the color indices, luminosity, de Vaucouleurs radius, inverse concentration index etc. [Karachentseva et al. (1994), Dobrycheva et al. (2012), Melnyk et al. (2012)]. We plotted the diagrams of color indices $g-i$ and one of the aforementioned parameters and discovered that these parameters may be used for galaxy classification into three classes: E – elliptical and lenticular, S – spirals Sa-Scd types, and L – late spirals Sd-Sdm and irregulars types. The accuracy is 98 % for E, 88 % for S, and 57 % for L types. The combinations of color indices $g-i$ and inverse concentration index $R50/R90$; color indices $g-i$ and absolute magnitude M_r gave the best result: 143 263 E type, 112 578 S type, 61 177 L type [Dobrycheva et al. (2017)].

Not enough classification accuracy for L type galaxies, we undergone the different MLM for providing a binary automated morphological classification. Why is binary one, because we decided compound S and L in one class of the Late type galaxies L. We tested different MLM for the SDSS DR9 samples of 317 018 galaxies at $z < 0.1$ using KNIME Analytic Platform and found that the Random Forest provides the highest accuracy, namely 91 % of galaxy types are classified correctly: 96 % Early (E) and 80 % Late (L) types [Dobrycheva et al. (2017)].

To improve accuracy of the results we developed own code for the automated morphological classification of galaxies and apply it to the sample of 60 561 galaxies from the SDSS DR9 at $0.02 < z < 0.06$. Results of this study are presented in this paper.

2. Galaxy sample from SDSS DR9

We used the sample of 60 561 galaxies from SDSS DR9 with the absolute magnitudes $-24^m < M_r < -19.4^m$ at $0.02 < z < 0.06$. As we said above the color indices, inverse concentration index $R50/R90$, and absolute magnitude are the good parameters for training MLM [Dobrycheva et al. (2015)][Dobrycheva et al. (2017)]. The absolute magnitude was obtained by the formula:

$$M_r = m_r - 5 \cdot \lg(D_L) - 25 - K_r(z) - ext_r,$$

where m_r - visual stellar magnitude in r band, D_L - distance luminosity, ext_r - the Galactic absorption in r, $K_r(z)$ - k-correction in r band according to [Chilingarian et al. (2010), Chilingarian et al. (2012)]. The color indices were obtained as:

$$M_g - M_i = (m_g - m_i) - (ext_g - ext_i) - (K_g(z) - K_i(z)),$$

where m_g and m_i – visual stellar magnitude in g and i band; ext_g and ext_i – the Galactic absorption

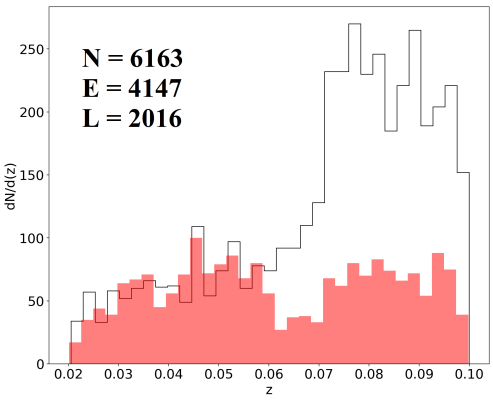

Figure 1: Distribution of galaxies by redshifts in the training sample (line - late type, pillars - early type galaxy)

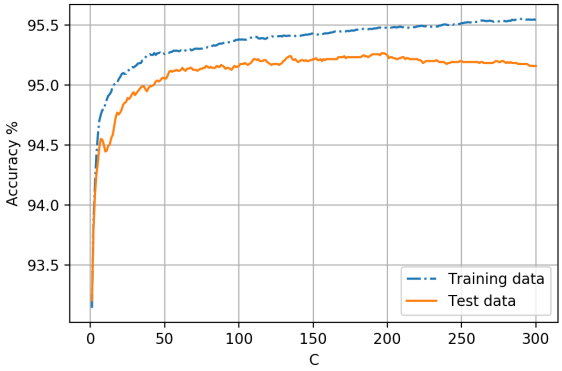

Figure 2: Training galaxy sample. Dependence of prediction accuracy for Logistic Regression Classifier on the "C" parameter

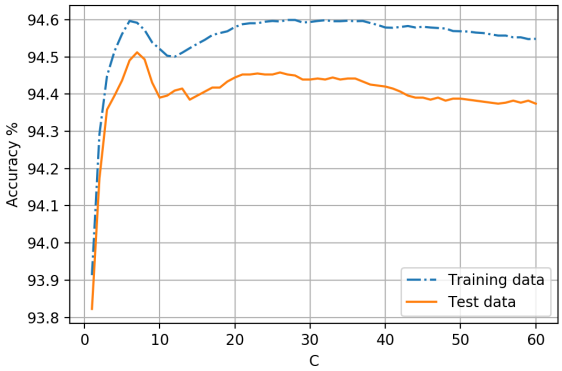

Figure 3: Training galaxy sample. Dependence of prediction accuracy for Support Vector Machine Classifier on the "C" parameter

in g and i band; $K_g(z)$ and $K_i(z)$ – k-correction in g and i band, respectively. Following the SDSS recommendation we involved limits $m_r < 17.7$ by visual magnitude in r-band to avoid typical statistical errors in spectroscopic flux.

3. Training galaxy sample

A training galaxy sample contains of 6 163 galaxies at $0.02 < z < 0.1$ with the absolute magnitudes $-24^m < M_r < -19.4^m$ from the SDSS DR9 (Fig. 1).

It was composed of several samples with certain morphological types of galaxies from our previous work and is based on the SDSS DR9. We split galaxies visually on two classes as E (including E, S0, S0a types) and L (from Sa to Irr types), which were selected randomly with different redshifts and luminosity. We collected 1) training sample, which contain 764 galaxies described by [Dobrycheva et al. (2018)], and 2) 5000 galaxies described by [Dobrycheva et al. (2015)]. After the beta run of machine learning for the unknown morphological types, we have done visual inspection of the randomly selecting galaxies and 3) added these galaxies to the training sample.

Additionally, we used an automatic regression method of discarding galaxies that strongly deviate from the mean value. We re-defined the mean of the magnitude in each filter and its scatter after each discard and then limited artificially this permissible deviation for the average. Thus, we got a sample of 6 163 galaxies.

4. Verification of Machine Learning methods for morphological classification

Logistic Regression is a statistical model that uses a logistic function to model the probability of a binary dependence of an object class on its features. The corresponding probability for each class may vary from 0 to 1 depending on the features [Tolles et al. (2016)]. A function that converts a logical factor into a probability is a logistic function. This can be extended to more than two classes. The model itself simply models the likelihood of output in terms of enter data and does not perform statistical classification (it is not a classifier). It includes the quantizer function and works by selecting the cutoff value and classifying the input as more likely than the cutoff as one class, below as another.

Support-Vector Machines (SVMs) are controlled learning models that analyze data used for classification and regression analysis. Given a set of classified training cutters, the SVM learning algorithm builds a model that determines the class of objects by their parameters [Smola et al. (2004)]. The SVM model is a representation of the points of the hyperspace objects so that the samples of the individual classes are separated

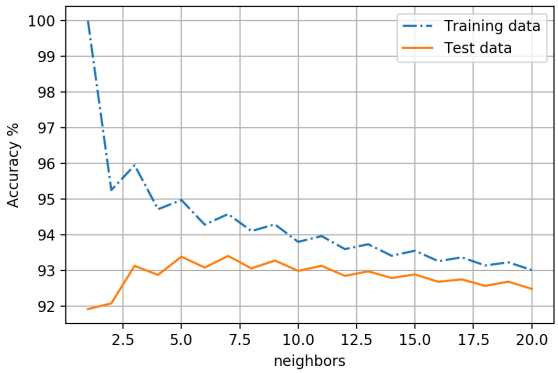

Figure 4: Training galaxy sample. Dependence of prediction accuracy for k-Nearest Neighbors Classifier on parameter "neighbors"

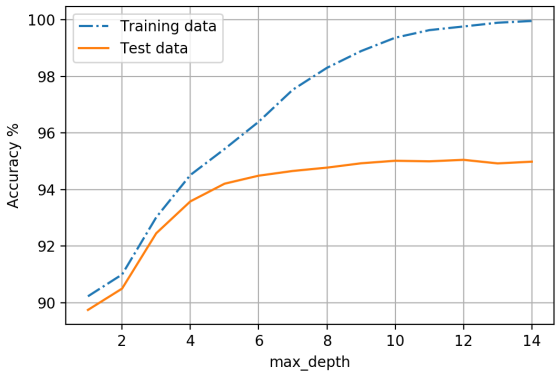

Figure 5: Training galaxy sample. Dependence of prediction accuracy for Random Forest Classifier on parameter "max_depth"

by a clear gap, which wide is maximized as possible. Then, the new objects are displayed in the same space and assumed to belong to the category based on the side of the gap to which they fall. This method has such disadvantages: it is used only for problems with two classes, it is impossible to calibrate the probability of getting to a certain class, the model parameters are difficult to interpret.

By default, Logistic Regression and SVC use L2 controller. For these methods, the "C" parameter determines the degree of regulation, where a higher degree of "C" corresponds to a lesser regulation. When "C" values height Logistic Regression and SVM trying to fit the models to the initial data as accurately as possible, then "C" is low models trying to look for a vector of coefficients closely to zero. Therefore, at low "C" values, the algorithm tries to fit the most data points, while at low "C" values it increases the contribution of each individual point. The accuracy dependencies of the "C" parameter for Logistic Regression and SVM are shown in Figures 2 and 3, respectively.

k-Nearest Neighbors (k-NN) is a simple non-parametric method used for classification. The object is classified by multiple votes of its neighbors, and the object is assigned to the class most common among its closest neighbors, where k-number of neighbors. If $k = 1$, then the object is simply assigned to the class of one nearest neighbor in the parameter space for the selected metric. Neighbors are drawn from the training dataset [Burkov et al. (2019)]. The accuracy dependencies of the number of neighbors for k-Nearest Neighbors is shown in Figure 4. A feature of the k-NN algorithm is related to its sensitivity to the local data structure. Advantages of this method: simple implementation, adaptation to the desired task by choosing a metric, interpret-ability. The disadvantages include: poor performance in tasks with many objects in the training sample; difficulties in finding the right weight and determining what features are required for classification; dependence on the selected metric.

Naive Bayes classifiers is a family of simple probability classifiers based on the application of Bayes' theorem with strong "naive" assumptions about independence between traits. This determines a certain statistical distribution of parameters for each of the classes [Soria et al. (2011)]. The probability of falling into a class depends on the ratio of the statistical density of the distribution of classes at a point for the selected model by independent parameters.The advantages of the method are as follows: high speed of work, easy interpret-ability of the results of the algorithm. The relatively low quality of classification and the inability to take into account the dependence of the classification result on a combination of features are the main disadvantages of this method. We used this method as a reference.

Random Forest is an ensemble classification and regression method that works by constructing a large number of decision trees and averaging the result of predicting individual trees. This helps to reduce the risk of overfiting. Tree models - where the target variable can take a discrete set of class values, are called classification trees [Burkov et al. (2019)]. In these trees, the leaves represent classes, and the branches represent the set of features that lead to these classes. In branching nodes there is a logical operator. The goal is to create a model that predicts the value of the target variable with the highest precision based on multiple input variables by calibrating the tree shape. Tree construction is recursive until the subset in the node has all the characteristics of the target variable, or when splitting the accuracy of the prediction will not remain constant. The accuracy dependencies of the maximum depth of the tree for Random Forest Classifier applied to the tra ining sample is shown in Figure 5. The main advantage of the method is the high productivity of training and forecasting; such decision trees can

be easily visualized and interpreted. The disadvantage is related to the method's propensity for retraining.

6. Results and Conclusion

Using own code in Scikit Learn Python[1] to predict correctly the galaxy morphology (Late and Early types) we verified several Machine Learning methods for binary morphological classification of galaxies. With this aim we used the sample of 60 561 galaxies from the SDSS DR9 survey with a redshift of $0.02 < z < 0.06$ and absolute magnitudes of $-24^m < M_r < -19.4^m$. Among the machine learning methods were as follows: Naive Bayes, Random Forest, Support Vector Machines, Logistic Regression, and k-Nearest Neighbor algorithm. To study the classifier, we used absolute magnitudes M_u, M_g, M_r, M_i, M_z, color indices $M_u - M_r, M_g - M_i, M_u - M_g, M_r - M_z$, and inverse concentration index to the center $R50/R90$.

Prediction accuracy was evaluated for each of these methods for training galaxy sample (see, Figures) and reaches the following values:

Naive Bayes Classifier – 0.886 (E – 0.920, L – 0.818) ± 0.01;

k-Nearest Neighbors Classifier – 0.945 (E – 0.9389, L – 0.958) ± 0.006;

Logistic Regression Classifier – 0.949 (E – 0.968, L – 0.911) ± 0.006;

Random Forest Classifier – 0.9545 (E – 0.967, L – 0.928) ± 0.003;

Support Vector Machine Classifier – 0.964 (E – 0.961, L – 0.969) ± 0.006.

All the above mentioned classifiers include the K-Fold Cross Validation method.

We compared these new results with previous one, which were made using the KNIME Analytics Platform 3.5.3 ([Dobrycheva et al. (2017)]). It turned out that the method of Random Forest and Support Vector Machine provide a highest accuracy (as in the previous study for the Random Forest Method), but with help of our code in Python we increased an accuracy from 92.9 % of correctly classified (96% – E and 84% – L) to 94,6% (96,9% – E and 89,7 % – L). The accuracy of the remaining methods also grew by 88% to 93% (see, Figure 6, where the images of the correct classification are presented).

So, using the Random Forest and Support Vector Machine Classifiers, and the data on color indices, absolute magnitudes, inverse concentration index of galaxies with visual morphological types, we were able to classify 60 561 galaxies from the SDSS DR9 with unknown morphological types and found 22 301 E and 38 260 L types among them. At the same time, the results of applying the Deep convolutional neural network (DL) to the images of redshift-limited ($z < 0.1$) sample of ∼ 300 000 galaxies from the SDSS DR9 by [Khramtsov et al. (2019)] with the same aim of binary morphological classification has been shown, for example, that DL method can classify rounded sources as Ellipticals but it can not catch the spectral energy distribution properties of galaxies more clearly than SVM, trained on the photometric features of galaxies.

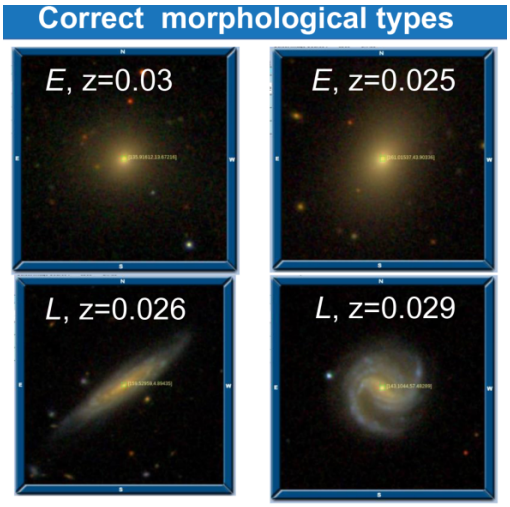

Figure 6: Images of galaxies from SDSS with the correctly classified morphology. Top: Early type (Ellipticals). Bottom: Late type (Spirals).

The problem points arise when we have cases of the face-on and edge-on galaxies (Figure 7). The most of these galaxies are miss-classified as Ellipticals (early type). The good case is that methods allow us to recover gravitational lenses (point-like sources, arcs) and the most of such miss-classifications are also among Ellipticals. So, we have overestimated number of Ellipticals and underestimated number of Spirals (about of 10 %). But this problem can be decided, when we will form training samples through several steps (pre-training, fine-tuning, and classification). The step of fine-tuning should include the limitations on the axes-ratio for Ellipticals, additional photometry parameters for the face-on galaxies, as well as trainings with images and spectral features of galaxies. Results of this approach as well as a conception of the automated morphological classification of a big data sample of galaxies with a wider redshift range will be given in other papers.

The Machine learning methods are an indispensable assistant in solving morphological classification since their first application to decide this problem with the ANN-algorithm [Storrie-Lombardi et al. (1992)]. They are also effective for reconstruction of Zone of Avoidance, distance modulus for local galaxies, gravitational lenses search, where the authors have own experience ([Vavilova et al. (2018)], [Elyiv et al. (2019)], [Sergeyev et al. (2018)], respectively).

[1] https://scikit-learn.org/

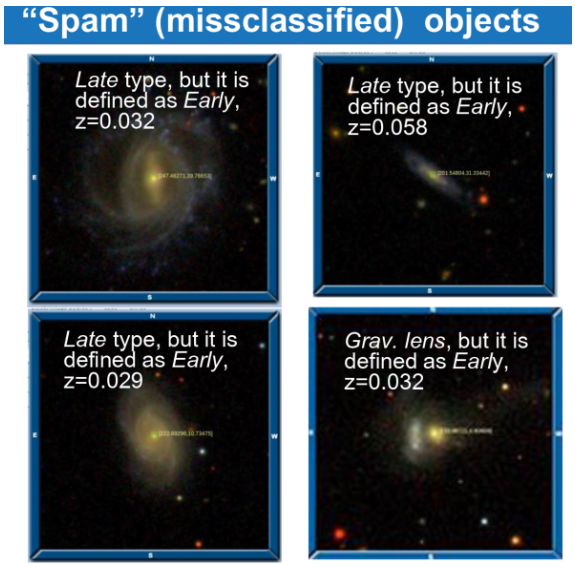

Figure 7: Images of galaxies from SDSS with the misclassified morphology. Top and Left Bottom: Late types (Spirals), which are classified as Early type (Ellipticals). Right Bottom: gravitational lens classified as Early type galaxy (Ellipticals).

Acknowledgements. This work was partially supported by the grant for Young Scientist's Research Laboratories (2018-2019, Dobrycheva D.V.) and the Youth Scientific Project (2019-2020, Dobrycheva D.V., Vasylenko M.Yu.) of the NAS of Ukraine.

References

Acero F., Ackermann M. et al.: 2015, *ApJS*, **218**, 41.

Al-Jarrah O. Y., Yoo P. D., Muhaidat S. et al.: 2015, Efficient machine learning for big data: A review. *Big Data Research*, **2(3)**, 87.

Andrae R., Melchior P. et al.: 2010, *A&A*, **522**, 19.

Banerji M., Lahav O. et al.: 2010, *MNRAS*, **406**, 342.

Barchi P. H., de Carvalho R. R., Rosa R. R. et al.: 2019, *arXiv:1901.07047*.

Blanton M. R., Bershady M. A., Abolfathi B. et al.: 2017, *ApJ*, **154**, 35.

Braude S.Ya., Rashkovsky S.L., Sidorchuk K.M. et al.: 2002, *Astrophysics and Space Science*, **280**, 235.

Burkov A.: 2019, The Hundred-Page Machine Learning Book.

Calderon V. F.; Berlind A. A.: 2019, *arXiv:1902.02680*.

Chilingarian I., Melchior A.L., Zolotukhin I.: 2010, *MNRAS*, **405**, 1409.

Chilingarian I., Zolotukhin I.: 2012, *MNRAS*, **419**, 1727.

Dobrycheva D., Melnyk O.: 2012, *AASP*, **2**, 42.

Dobrycheva D.V.: 2013, *OAP*, **26**, 187.

Dobrycheva D. V., Melnyk O. V., Vavilova I. B. et al.: 2015, *Astrophysics*, **58**, 168.

Dobrycheva D. V. et al.: 2017, *arXiv:1712.08955*.

Dobrycheva D.V., Vavilova I. B., Melnyk O. V. et al.: 2018, *Kinemat. Phys. Celest. Bodies*, **34**, 290.

Dominguez S.H.; Huertas-Company M., Bernardi M. et al.: 2018, *MNRAS*, **476**, 3661.

Elyiv A., Melnyk O. et al.: 2009, *MNRAS*, **394**, 1409.

Elyiv A.A. et al.: 2019, *arXiv:1910.07317*.

Gunn J.E., Carr M. et al.: 1998, *ApJ*, **116**, 3040.

Ivezic Z., Connelly A.J., VanderPlas J.T. et al.: 2014, Statistics, Data Mining, and Machine Learning in Astronomy, by Z. Ivencić et al. Princeton, NJ: Princeton University Press.

Karachentseva V.E. et al.: 1994, *Bull. SAO*, **37**, 98.

Khramtsov V., Sergeyev A., Spiniello, C. et al.: 2019, *arXiv:1906.01638*.

Khramtsov V. et al.: 2019, *OAP this issue*.

Lee J.C., Gil de Paz A. et al.: 2011, *ApJS*, **192**, 33.

Melnyk O.V., Dobrycheva D.V., Vavilova I.B.: 2012, *Astrophysics*, **55**, 293.

Norris R.P.: 2017, *Nature Astronomy*, **1**, 671.

Pierre M., Pacaud F. et al.: 2016, *A&A*, **592**, 16.

Rosen S.R., Webb N.A. et al.: 2016, *A&A*, **590**, 22.

Scoville N., Abraham R.G. et al.: 2007, *ApJS*, **172**, 38.

Sergeyev A., Spiniello C. et al.: 2018, *AAS*, **2**, 189.

Skrutskie M.F., Cutri R.M., Stiening R. et al.: 2006, *Astron. J.*, **131**, 1163.

Soria D., Garibaldi J. M., Ambrogi F, et al.: 2011, *Knowledge Based Systems*, **24**, 775.

Smola A. J., Scholkopf B.: 2004, *Statistics and Computing*, **14**, 199.

Srivastava A. N. (Ed.).: 2012, Advances in machine learning and data mining for astronomy. Chapman and Hall/CRC.

Storrie-Lombardi M.C., Lahav O., Sodre L.Jr. et al.: 1992, *MNRAS*, **259**, 8.

Tolles J., Meurer W. J: 2016, Logistic Regression Relating Patient Characteristics to Outcomes, ISSN 0098-7484

Vavilova I.B., Karachentseva V.E., Makarov D.I. et al.: 2005, *Kinemat. Physics Celest. Bodies*, **21**, 3.

Vavilova I.B., Melnyk O.V., Elyiv A.A.: 2009, *Astron. Nachr.*, **330**, 1004.

Vavilova I.B.: 2016, *OAP*, **29**, 109.

Vavilova I.B., Elyiv A.A., Vasylenko M.Yu.: 2018, *Radio Phys. Radio Astron.*, **23**, 244.

Voges W., Aschenbach B., Boller Th. et al.: 2000, *VizieR On-line Data Catalog*, **IX/29**.

Wright E. L., Eisenhardt P. R. M., Mainzer A. K.: 2010, *ApJ*, **140**, 1868.

Zaane O. R.: 1999. Introduction to data mining.

Zhixian Ma et al.: 2018 *arXiv:1812.07190*.