Use of Machine Learning for Anomaly Detection Problem in Large Astronomical Databases

Konstantin Malanchev[1,2,*], Alina Volnova[3], Matwey Kornilov[1,2,+], Maria Pruzhinskaya[1,++], Emille Ishida[4], Florian Mondon[4], and Vladimir Korolev[5,6]

[1] Lomonosov Moscow State University, Sternberg Astronomical Institute, Universitetsky pr. 13, Moscow, 119234, Russia
* malanchev@sai.msu.ru
+ matwey@sai.msu.ru
++ pruzhinskaya@gmail.com

[2] National Research University Higher School of Economics, 21/4 Staraya Basmannaya Ulitsa, Moscow, 105066, Russia
[3] Space Research Institute of the Russian Academy of Sciences (IKI), 84/32 Profsoyuznaya Street, Moscow, 117997, Russia
[4] Université Clermont Auvergne, CNRS/IN2P3, LPC, F-63000 Clermont-Ferrand, France
[5] Central Aerohydrodynamic Institute, 1 Zhukovsky st, Zhukovsky, Moscow Region, 140180, Russia
[6] Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia

Abstract. In this work, we address the problem of anomaly detection in large astronomical databases by machine learning methods. The importance of such study is justified by the presence of a large amount of astronomical data that cannot be processed only by human resource. We focus our attention on finding anomalous light curves in the Open Supernova Catalog. Few types of anomalies are considered: the artifacts in the data, the cases of misclassification and the presence of previously unclassified objects. On a dataset of ~ 2000 supernova (SN) candidates, we found several interesting anomalies: one active galactic nucleus (SN2006kg), one binary microlensing event (Gaia16aye), representatives of rare classes of SNe such as super-luminous supernovae, and highly reddened objects.

Keywords: Machine learning; Isolation forest; Gaussian processes; Supernovae; Transients

1 Introduction

During the last couple of decades, astronomy eventually became the source of huge amounts of data produced by different dedicated surveys and experiments, which require careful processing to extract valuable information. Gigabytes of data are collected daily in every domain of electromagnetic spectrum: in high-energy range [1], optics [2, 3], and radio [4], as well as in cosmic particles window [5] and gravitational waves [6, 7]. The search for yet unknown statistically significant features of astronomical objects,

as well as the distinction of real features from processing artifacts is an important problem of the automated data analysis.

Supernovae (SNe) are among the most numerous objects discovered in astronomy, and their total amount increases by several thousand per year. These objects help to solve many astronomical puzzles: they produce the majority of heavy chemical elements [8] and high energy cosmic rays [9], they trigger star formation in galaxies [10]. Moreover, the study of different types of SNe allows us to probe the composition and distance scale of the Universe imposing strong constraints on the standard cosmological model [11]. In the last 15 years several large surveys have already gathered many observational data on SNe and their candidates (Carnegie Supernova Project – CSP [12], the Panoramic Survey Telescope and Rapid Response System – Pan-STARRS [13] and the Dark Energy Survey – DES [14]). The surveys of new generation, like Large Synoptic Survey Telescope [15], will produce data of unprecedented volume and complexity.

The exponential growth of astronomical data volume makes the use of machine learning (ML) methods inevitable in this field [16]. Most of the ML efforts in astronomy are concentrated in classification [e.g., 17] and regression [e.g., 18] tasks. A large variety of ML methods were applied to supervised photometric SN classification problem [19–21] and unsupervised characterization from spectroscopic observation [e.g., 22].

Astronomical anomaly detection is the field where ML methods may be used quite effectively taking into account the enormous amount of data that has been gathered, however, they have not been fully implemented yet. Barring a few exceptions, most of the previous studies may be divided into only two different trends: clustering [23] and subspace analysis [24] methods. More recently, random forest algorithms have been used extensively by themselves [25] or in hybrid statistical analysis [26]. Although all of this has been done to periodic variables there is not much done for transients and even less for supernovae.

Supernovae surveys detect hundreds of SNe candidates per year, but the lack of spectroscopic information makes the processing algorithms to classify discovered SNe basing on secondary features (proximity to the galaxy, monotonous flux changing with time, absolute magnitude, etc.) Anomaly detection may solve two problems: (a) minimize the contamination of non-SNe in large supernova databases, and (b) find inside the SNe data rare or new classes of objects with unusual properties.

In this paper, we suggest the algorithm of anomaly detection using the isolation forest method and basing on real photometrical data from the Open Supernova Catalog (OSC) [27, 28].

2 Data Preprocessing

2.1 Open Supernova Catalog

The Open Supernova Catalog [27] is constructed by combining many publicly available data sources. It includes many catalogs and surveys, such as Pan-STARRS [29], the SDSS Supernova Survey [30], the All-Sky Automated Survey for Supernovae (ASAS-SN [31]), the intermediate Palomar Transient Factory (iPTF [32]) among others, as well as information from individual studies. It represents an open repository for supernova metadata, light curves, and spectra in an easily downloadable format. This catalog also includes some contamination from non-SN objects. It contains data for more than 55000 SNe candidates among which ~13000 objects have >10 photometric observations and for ~7500 spectra are available.

The catalog stores the light curves (LCs) data in different magnitude systems. Since we need a homogeneous data sample, we extracted only the LCs in *BRI*, *gri,* and *g'r'i'* filters. We assume, that g'r'i' filters are close enough to gri filters to consider them as the same filters. We also transform BRI magnitudes to gri using the Lupton's photometrical equations [33]. We also require a minimum of three photometric points in each filter with 3-day binning. After this first cut, our sample consists of ~3000 objects.

2.2 Light Curves Approximation

Traditionally, ML algorithms require a homogeneous input data matrix which, unfortunately, is not the case with supernovae. A commonly used technique to transform unevenly distributed data into a uniform grid is to approximate them with Gaussian processes (GP [34]). Usually, each light curve is approximated by GP independently. However, in this study we use a Multivariate Gaussian Process [35] approximation. For each object it takes into account the correlation between light curves in different bands, approximating the data by GP in all filters in a one global fit (for details see Kornilov et al. 2019, in prep.). As an approximation range we chose [–20; +100] days. We also extrapolated the GP approximation to fill this range if needed. With this technique we can reconstruct the missing parts of LC from its behavior in other filters.

Gaussian process is based on the so-called kernel, a function describing the covariance between two observations. The kernel used in our implementation of Multivariate Gaussian Process is composed of three radial-basis functions

$$k(t_1, t_2) = e^{\left(\frac{-(t_2 - t_1)^2}{2l_i^2}\right)},$$

where i denotes the photometric band, and l_i are the parameters of Gaussian process to be found from the light curve approximation. In addition, Multivariate Gaussian Process kernel includes 6 constants, three of which are unit variances of basis processes and three others describe their pairwise correlations. Totally, Multivariate Gaussian Process has 9 parameters to be fitted.

Once the Multivariate Gaussian Process approximation was done, we visually inspected the resulting light curves. Those SNe with unsatisfactory approximation were removed from the sample (mainly the objects with bad photometric quality). Since each

object has its own flux scale due to the different origin and different distance, we normalized the flux vector by its maximum value. Based on the results of this approximation, for each object we extracted the kernel parameters, the log-likelihood of the fit, LC maximum and normalized photometry in the range of [−20, +100] days with 1-day interval relative to the maximum. These values were used as features for the ML algorithm (Sect. 3). Our final sample consists of 1999 objects, ∼30% of which have at least one spectrum in the OSC. Less than 5% of our sample have <20 photometric points in all three filters.

2.3 Dimensionality Reduction

After the approximation procedure, each object has 374 features: 121×3 normalized fluxes, the LC flux maximum, 9 fitted parameters of the Gaussian process kernel, and the log-likelihood of the fit.

We apply the anomaly detection algorithm not only to the full data set but also to the dimensionality-reduced data. The reason for this is that the initial high dimensional feature space can be too sparse for the successful performance of the isolation forest algorithm. We applied t-SNE [36], a variation of the stochastic neighbor embedding method [37], for the dimensionality reduction of the data. As a result of the dimensionality reduction, we obtain 8 separate reduced data sets corresponding to 2 to 9 t-SNE features (dimensions).

3 Anomaly Detection

3.1 Isolation Forest

Isolation forest [38, 39] is an outlier detection technique that identifies anomalies instead of normal observations. It is built on an ensemble of random isolation trees. Each isolation tree is a space-partitioning tree similar to the widely known Kd-tree [40]. However, in contrast to the Kd-tree, a space coordinate (a feature) and a split value are selected at random for every node of the isolation tree. The tree is built until each object of a sample is isolated in a separate leaf – the shorter path corresponds to a higher anomaly score. For each object, the measure of its normality is the arithmetic average of the depths of the leaves into which it is isolated. The idea of identifying normal data vs. anomalies is presented in Fig. 1.
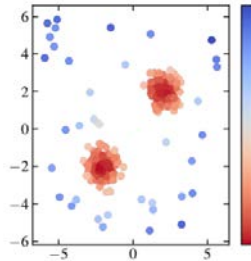
Fig. 1. This `scikit-learn` example presents the generated 2D dataset. Regions of high density are normal data while the outliers are spread around, which is also illustrated by the colour – bluer colour means more anomalous behaviour
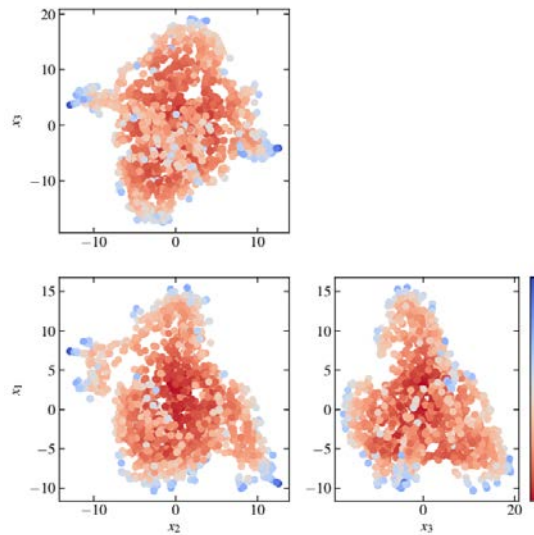

Fig. 2. Three-dimensional t-SNE reduced data after application of the isolation forest algorithm. Each point represents a supernova light curve from the data set projected into the three-dimensional space with the coordinates (x1, x2, x3). The intensity of the colour indicates the anomaly score for each object as estimated by the isolation forest algorithm – bluer colour means more anomalous light curve behaviour

We run the isolation forest algorithm on 10 data sets:

I. data set of 364 photometric characteristics (121×3 normalized fluxes, the LC flux maximum),

II. data set of 10 parameters of the Gaussian process (9 fitted parameters of the kernel, the log-likelihood of the fit),

III. 8 data sets obtained by reducing 374 features to 2–9 t-SNE dimensions (Sect. 2.3).

For each data set we obtained a list of anomalies. Contamination levels were set to 1% (20 objects with highest anomaly score) for data sets I and II. For all data sets in case III we considered 2% contamination (40 objects with highest anomaly score). This larger contamination was chosen to take into account the influence of the dimensionality reduction step in the final data configuration. Given different representations of the data and the stochastic nature of the isolation forest algorithm, the same object can be assigned a different anomaly score depending on how many t-SNE dimensions are used. Thus, only those objects which were listed within the 2% contamination in at least 2 of the data sets in case III were included in the resulting group of objects to further astrophysical analysis.

An example of the isolation forest algorithm applied to the three-dimensional reduced data set is shown in Fig. 3.

3.2   Evaluation of t-SNE Technique

Since t-SNE is a stochastic technique, we have also taken additional precautions to ensure that the resulting anomaly list does not depend on the t-SNE initial random state. For each number of dimensions, we run the t-SNE 1000 times. Then, the isolation forest algorithm is applied to the newly obtained reduced dataset and a list of anomalies is produced. Next, we counted how often each supernova is listed in the anomaly list. Fig. 3 shows the distribution of supernovae by the frequency of appearance in anomaly list for the three-dimensional t-SNE reduced dataset (filled bars). The y-axis is normalized to the total number of runs.
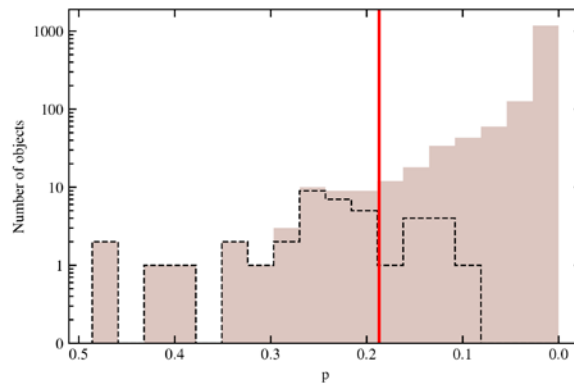

Fig. 3. The distribution of supernovae by the frequency of appearance in anomaly list for the three-dimensional t-SNE reduced dataset (filled bars). The red line denotes 2% of supernovae with the highest anomaly score contamination (40 objects). The distribution of supernovae that were subjected to the further analysis as anomalies in this work are marked by dashed line. The y-axis is normalized to the total number of runs (i.e., 1000)

4 Results

The isolation forest algorithm found ~100 anomalies among 1999 objects from all our samples. We inspected visually the LCs of selected anomalies and analyzed them using other publicly available information. Basing on this analysis, we decided whether the object is an anomaly or not. Among the detected anomalies, there are few cases of miss-classifications, representatives of rare classes of SNe and highly reddened objects. Here we list a few particular cases.

4.1 Peculiar Supernovae

By their spectral and photometric characteristics, the "normal" supernovae are historically divided into two wide types: Type I and Type II. The more recent classification distinguishes Type Ia, Ib, Ic, IIn, IIb, IIP, IIL supernovae. In terms of physics Type Ia SNe are a thermonuclear explosion of a white dwarf which mass exceeded the Chandrasekhar limit either due to accretion from a companion star or by a merging of two white dwarfs [41–43]. These SNe tend to have approximately the same luminosity in maximum and are considered as standard candles for cosmological scale estimates [44, 45]. However, the class of SNe Ia is not homogeneous – some of SNe are on average by 0.2–0.3 magnitudes brighter in maximum than others and some of them are on the contrary subluminous and fast-declining [46, 47]. The presence of non-standard SNe Ia in cosmological samples may introduce a systematic bias [48].

Other types of SNe mark the death of massive stars during the collapse of the core. The envelopes of these stars extend to hundreds of solar radii and contain large amounts of hydrogen (Type II). More massive progenitors of core-collapse supernovae can lose mass by the stellar wind and end their lives losing all (Ib) or part of the hydrogen envelope (IIb). An even more effective stellar wind can blow out not only the hydrogen but also the helium envelope (Ic).

SN2013cv is a peculiar Type Ia supernova with a large peak optical and UV luminosity and with an absence of iron absorption lines in the early spectra. It was suggested [49] to be an intermediate case between the normal and super-Chandrasekhar events.

SN2016bln/iPTF16abc belongs to a subtype of over-luminous Type Ia SNe. Its early-time observations show a peculiar rise time, non-evolving blue colour, and unusual strong C II absorption. These features can be explained by the ejecta interaction with nearby, unbound material or/and significant ^{56}Ni mixing within the SN ejecta [50].

SN2016ija was first suggested to be an early time 91T-like SN Ia with few features and red continuum. It has been also associated to the outburst in an obscured luminous blue variable, an intermediate luminosity red transient or a luminous red nova [51]. The subsequent spectroscopic follow-up revealed broad H$_\alpha$ and calcium features, leading to a classification as a highly extinguished Type II supernova [52].

4.2 Superluminous SNe

Superluminous SNe are supernovae with an absolute peak magnitude M<–21 mag in any band. According to [53] SLSN can be divided into three broad classes: SLSN-I without hydrogen in their spectra, hydrogen-rich SLSN-II that often show signs of interaction with CSM, and finally, SLSN-R, a rare class of hydrogen-poor events with slowly evolving LCs, powered by the radioactive decay of ^{56}Ni. SLSN-R are suspected to be pair-instability supernovae: the deaths of stars with initial masses between 140 and 260 solar masses. Our isolation forest algorithm found 4 SLSNe in our samples.

4.3 Misclassifications

SN2006kg was first classified as a possible Type II SN [54]. It is also appeared as Type II spectroscopically confirmed supernova in table 6 of [55]. However, further analysis of 3.6-m New Technology Telescope spectrum revealed that SN2006kg is an active galactic nucleus [56, 30].

Gaia16aye is an object with the most non-SN-like behavior among our set of outliers. In [57] it was reported to be a binary microlensing event – gravitational microlensing of binary systems – the first ever discovered towards the Galactic Plane.

Our analysis also revealed that 16 of detected anomalies (all from the SDSS SN candidate catalog [30]) are likely to be stars or quasars. First, we did not find any signature of supernovae on the corresponding multicolour light curves. Second, according to SDSS DR15 [58], 10 of these objects are denoted as STAR. The other 6 objects have a BOSS [59] spectrum with class "QSO" and have high redshifts.

More detailed analysis of the detected anomalies is presented in [60].

5 Conclusions

The amount of astronomical data increases dramatically with time and is already beyond human capabilities. The astronomical community already has dozens of thousands of SN candidates, and LSST survey will discover over ten million supernovae in the forthcoming decade [61]. Only a small fraction of them will receive a spectroscopic confirmation. This motivates a considerable effort in photometric classification of supernovae by types using machine learning algorithms. There is, however, another aspect of the problem: any large photometric SN database would suffer from the non-SN contamination (novae, kilonovae, GRB afterglows, AGNs, etc.). Moreover, the database will inevitably contain the astronomical objects with unusual physical properties – anomalies. In this study, we show that the isolation forest algorithm may be rather efficient in solving this problem. This method identified ~100 potentially interesting objects from 1999 supernova candidates extracted from the Open Supernova Catalog, ~30% of which were confirmed to be non-SN events or representatives of the rare SN classes. Among these objects, we report for the first time the 16 star/quasar-like objects misclassified as SNe.

It is important to note that these results are not expected to be complete. There are several known SLSNe in our sample, which were not identified as anomalies, and sev-

eral objects with very distinguishing features, which do not affect the LC shape significantly, so the algorithm missed them. This may indicate some defects in Gaussian Processes approximation of initial observed LCs. Nevertheless, the above results provide clear evidence of the effectiveness of automated anomaly detection algorithms for photometric SN light curve analysis. This approach may be crucial for future surveys, like LSST, when the enormous amount of data make the search of outliers impossible for human abilities.

The code of this work and the data are available at http://xray.sai.msu.ru/snad/.

Acknowledgements

References

1. GBM/Fermi Homepage, https://fermi.gsfc.nasa.gov/science/instruments/gbm.html
2. SDSS-DR12 Homepage, https://www.sdss.org/
3. Gaia project Homepage, http://sci.esa.int/gaia/
4. Event Horizon Telescope Homepage, https://eventhorizontelescope.org/
5. IceCUBE observatory Homepage, https://icecube.wisc.edu/
6. LIGO Homepage, https://www.ligo.caltech.edu/
7. Virgo observatory Homepage, www.virgo-gw.eu/
8. Nomoto, K., Kobayashi, C., and Tominaga, N.: Nucleosynthesis in Stars and the Chemical Enrichment of Galaxies. ARA&A **51**, 457 (2013).
9. Morlino, G.: High-energy cosmic rays from Supernovae. In Handbook of Supernovae, ed. Athem W. Alsabti and Paul Murdin, 1711 p. (2017).
10. Chiaki, G., Yoshida, N., and Kitayama, T.: Low-mass star formation triggered by early Supernova explosions. ApJ **762**, 50 (2013).
11. Scolnic, D.M., Jones, D.O., Rest, A., et al.: The complete light-curve sample of spectroscopically confirmed SNe Ia from Pan-STARRS1 and cosmological constraints from the combined pantheon sample. ApJ **859**, 101 (2018).
12. Carnegie Supernova Project Homepage, https://csp.obs.carnegiescience.edu/
13. Panoramic Survey Telescope and Rapid Response System Homepage, https://pan-starrs.stsci.edu/
14. Dark Energy Survey Homepage, https://www.darkenergysurvey.org/

15. Large Synoptic Survey Telescope Homepage, https://www.lsst.org/

16. Ball, N.M. and Brunner, R.J.: Data mining and machine learning in astronomy. International Journal of Modern Physics D **19**, 1049 (2010).

17. Ishida, E.E.O., Beck, R., González-Gaitán, S., et al.: Optimizing spectroscopic follow-up strategies for supernova photometric classification with active learning. MNRAS **483**, 2 (2019).

18. Beck, R., Lin, C.A., Ishida, E.E.O., et al.: On the realistic validation of photometric redshifts. MNRAS **468**, 4323 (2017).

19. Brunel, A., Pasquet, J., Pasquet, J., et al.: A CNN adapted to time series for the classification of Supernovae. arXiv e-prints, p. arXiv:1901.00461 (2019).

20. Pasquet, J., Pasquet, J., Chaumont, M., and Fouchez D.: PELICAN: deeP architecturE for the LIght Curve ANalysis. arXiv e-prints, p. arXiv:1901.01298 (2019).

21. Möller, A. and de Boissière, T.: SuperNNova: an open-source framework for Bayesian, Neural Network based supernova classification. arXiv e-prints, p. arXiv:1901.06384 (2019).

22. Muthukrishna, D., Parkinson, D., and Tucker, B.: DASH: Deep learning for the automated spectral classification of Supernovae and their Hosts. arXiv e-prints, p. arXiv:1903.02557 (2019).

23. Rebbapragada, U., Protopapas, P., Brodley, C.E., and Alcock, C.: Finding anomalous peri-odic time series. Machine Learning **74**, 281 (2009).

24. Henrion, M., Hand, D.J., Gandy, A., Mortlock, D.J.: CASOS: a subspace method for anom-aly detection in high dimensional astronomical databases. Statistical Analysis and Data Min-ing: The ASA Data Science Journal **6**, 53 (2013).

25. Baron, D., Poznanski, D.: The weirdest SDSS galaxies: results from an outlier detection algorithm. MNRAS **465**, 4530 (2017).

26. Nun, I., Pichara, K., Protopapas, P., and Kim, D.-W.: Supervised detection of anomalous light curves in massive astronomical catalogs. ApJ **793**, 23 (2014).

27. Guillochon, J., Parrent, J., Kelley, L.Z., and Margutti, R.: An open catalog for Supernova Data. ApJ **835**, 64 (2017).

28. Open Supernova Catalog Homepage, https://sne.space/

29. Chambers, K.C., Magnier, E.A., Metcalfe, N., et al.: The Pan-STARRS1 Surveys. arXiv e-prints, arXiv:1612.05560 (2016).

30. Sako, M., Bassett, B., Becker, A.C., et al.: The data release of the Sloan Digital Sky Survey-II Supernova Survey. PASP **130**, 064002 (2018).

31. Holoien, T.W.-S., Brown, J.S., Vallely, P.J., et al.: The ASAS-SN bright supernova cata-logue-IV. 2017. MNRAS **484**, 1899 (2019).

32. Cao, Y., Nugent, P. E., and Kasliwal, M.M.: Intermediate palomar transient factory: realtime image subtraction pipeline. PASP **128**, 114502 (2016).

33. Lupton's transformation equations for SDSS, http://www.sdss3.org/dr8/algorithms/sdssUBVRITransform.php

34. Rasmussen, C.E. and Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press (2005).

35. Multivariate Gaussian Processes code, http://gp.snad.space/en/latest

36. Maaten, L. v. d. and Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579 (2008).

37. Hinton, G.E. and Roweis, S.T.: Advances in neural information processing systems, 857–864 (2003).

38. Liu, F.T., Ting, K.M., and Zhou, Z.-H.: In 2008 Eighth IEEE International Conference on Data Mining, 413–422 (2008).

39. Liu, F.T., Ting, K.M., and Zhou, Z.-H.: Isolation-based Anomaly Detection. ACM Trans. Knowl. Discov. Data **6**, 1 (2012).

40. Bentley, J.L.: Multi dimensional binary search trees used associative searching. Commun. ACM **18**, 509 (1975).

41. Whelan, J. and Iben, Jr. I.: Binaries and Supernovae of type I. ApJ **186**, 1007 (1973).

42. Iben Jr.I. and Tutukov A.V.: Supernovae of type I as end products of the evolution of bina-ries with components of moderate initial mass (M not greater than about 9 solar masses). ApJS **54**, 335 (1984).

43. Webbink, R.F.: Double white dwarfs as progenitors of R Coronae Borealis stars and Type I supernovae. ApJ **277**, 355 (1984).

44. Perlmutter, S., Aldering, G., Goldhaber, G., et al.: Measurements of Ω and Λ from 42 High-Redshift Supernovae. ApJ **517**, 565 (1999).

45. Riess, A.G., Filippenko, A.V., Challis, P., et al.: Observational evidence from Supernovae for an accelerating universe and a cosmological constant. AJ **116**, 1009 (1998).

46. Filippenko, A.V., Richmond, M.W., Branch, D., et al.: The subluminous, spectroscopically peculiar type IA supernova 1991bg in the elliptical galaxy NGC 4374. AJ **104**, 1543 (1992).

47. Filippenko, A.V., Richmond, M.W., Matheson, T., et al.: The peculiar Type IA SN 1991T - Detonation of a white dwarf? ApJ **384**, L15 (1992).

48. Scalzo R., Aldering, G., Antilogus, P., et al.: A Search for new candidate Super-Chandra-sekhar-mass Type Ia Supernovae in the Nearby Supernova Factory Data Set. ApJ **757**, 12 (2012).

49. Cao Y., Johansson, J., Nugent, P. E., et al.: Absence of Fast-moving Iron in an Intermediate Type Ia Supernova between normal and super-chandrasekhar. The Astrophysical Journal **823**, 147 (2016).

50. Miller, A.A., Cao, Y., Piro, A.L., et al.: Early Observations of the Type Ia Supernova iPTF 16abc: A Case of Interaction with Nearby, Unbound Material and/or Strong Ejecta Mixing. ApJ **852**, 100 (2018).

51. Blagorodnova N., Neill, J.D., Kasliwal, M., et al.: Follow-up observations of DLT16am/AT2016ija with SEDM. The Astronomer's Telegram, 9787 (2016).

52. Tartaglia, L., Sand, D., Valenti, S., et al.: The Early Detection and Follow-up of the Highly Obscured Type II Supernova 2016ija/DLT16am. ApJ **853**, 62 (2018).

53. Gal-Yam, A.: Luminous Supernovae. Science **337**, 927 (2012).

54. Bassett, B., Becker, A., Brewington, H., et al.: SUPERNOVAE 2006kg-2006lc. Central Bureau Electronic Telegrams **688**, 1 (2006).

55. Sako M., Bassett, B., Becker, A., et al.: The Sloan Digital Sky Survey-II Supernova Survey: Search Algorithm and Follow-up Observations. AJ **135**, 348 (2008).

56. Östman L., Nordin, J., Goobar, A., et al.: NTT and NOT spectroscopy of SDSS-II superno-vae. A&A **526**, A28 (2011).

57. Wyrzykowski L., Leto, G., Altavilla, G., et al.: Gaia16aye is a binary microlensing event and is crossing the caustic again. The Astronomer's Telegram, 9507 (2016).

58. SDSS-DR15 Data, http://skyserver.sdss.org/dr15/en/tools/explore/summary.aspx

59. Smee, S.A., Gunn, J.E., Uomoto, A., et al.: The Multi-object, Fiber-fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey. AJ **146**, 32 (2013).

60. Pruzhinskaya, M.V., Malanchev, K.L., Kornilov, M.V., et al.: Anomaly Detection in the Open Supernova Catalog. MNRAS, Volume 489, Issue 3, Pages 3591-3608 (2019).

61. LSST Science Collaboration et al.: LSST Science Book, Version 2.0. arXiv e-prints, arXiv:0912.0201 (2019).

62. van der Walt, S., Colbert, S.C., and Varoquaux, G.: The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science and Engineering **13**, 22 (2011).
63. Hunter, J.D.: Matplotlib: A 2D Graphics Environment. Computing in Science and Engineering **9**, 90 (2007).
64. Jones, E., Oliphant, T., Peterson, P., et al., SciPy: Open source scientific tools for Python, http://www.scipy.org/ (2001)
65. McKinney, W.: Data Structures for Statistical Computing in Python. In: van der Walt S., Millman, J. (eds.), Proceedings of the 9th Python in Science Conference, 51–56 (2010).
66. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825 (2011).