Implementation of machine learning algorithms in the Sloan Digital Sky Survey DR14 analysis

D A Petrusevich

Russian Technological University (MIREA), Prospekt Vernadskogo, 78, Moscow, 119454, Russia

E-mail: petrdenis@mail.ru

Abstract. The fourth edition of the Sloan Digital Sky Survey has been investigated in the paper. There are a few telescopes analyzing sky at different frequencies. They generate a lot of statistical data combined into datasets. One of them is explored in the paper. The handled dataset contains information about three types of objects: stars, quasars and galaxies. Efforts of physicists aren’t enough to investigate vast amount of data. The goal of machine learning implemented in this area is to solve the most tasks of classification in automatical way. Attention should be paid only to some complicated cases. Information in such datasets is already marked up in order to apply classification algorithms and models. Review of literature has shown that neural networks are often used to investigate such datasets that could be handled with simple models. In this research some simple classification models are implemented, as well there are results of ensemble algorithms implementation. Advantages and disadvantages of their implementations are described, physical explanation of classifiers’ structure is presented when it’s possible. Results and conclusions could be used in processing of other astronomical datasets.

1. Introduction

Today data analysis and machine learning are implemented in a lot of different domains of knowledge.

Astronomical objects investigation was based on the changes of light brightness, on movement fixation at the consequential photos, on the spectral analysis and etc. Such data processing was made manually. Now rapid development of machine learning allows physicists use data analysis methods in their domain. There’s vast amount of statistical data obtained with telescopes. Individual scientists aren’t able to handle this flow. Search of astronomical objects with specific properties is transformed into statistical task. Data scientists can participate in these problems’ solutions by means of astronomical datasets analysis.

Raising computing power of CPU and GPU has made possible to construct more exact and thorough numerical solutions of well-known gravitational tasks, for example: gravitational structure of our galaxy, dynamics of objects near the Sagittarius A* black hole (center of Milky Way galaxy) [1], solutions of system “planet-satellite” dynamics with higher precision than in case of classical solutions [2]. Recently a lot of attention is paid to the problem of gravitational waves analysis and bringing information from electromagnetic and gravitational fields together [3]. A set of problems that could be solved with machine learning methods include 1) star type classification (pulsar search, distinguishing pulsar signals from noise [4-7], 2) Ia type supernova search (which are used to measure distance to other galaxies), 3) exoplanet search and etc.
From the informational technologies point of view astronomical data analysis tasks are divided into two main groups. If handled data array doesn’t include graphical information it can be processed with mathematical statistics and data science methods \[4, 6, 8-12\]. Datasets containing sky images are analyzed with methods of computer graphics, spectral analysis and computer vision \[6, 7, 13-15\].

Still the presented papers on the statistical astronomical data processing contain only the simplest models (decision trees, random forest and sometimes logistic regression and simple neural networks). This paper includes computational experiment established to implement a set of popular data science models and a set of complex methods (for example, boosting and ensembles of classifiers). More complex classification models based on genetic algorithms \[16\] or neural networks are also implemented but it’s clear from the results of this research that they’re too complicated for these tasks \[5, 6\].

2. The dataset structure and classification quality metrics
Sloan Digital Sky Survey DR14 \[17\] is a dataset including 10000 records. Every object has got 17 parameters and one class label (star, quasar or galaxy). There are 4152 stars, 4998 galaxies and 850 quasars. Most of parameters are just spatial coordinates that can be excluded in the classification problem solution. There are 6 characteristics left and one column describing object’s type.

Objects are described with 6 values: 5-color broadband photometric system \{u, g, r, i, z\} (star magnitude in each filter) and object’s red shift (parameter redshift that means wavelength shift to the red part of spectrum for heading away objects).

Correlation between characteristics is seen in the table 1. Diagonal elements are correlations between parameters and themselves. They are equal to 1. The accuracy is two decimal places. All values about classifiers’ quality are rounded to three places to the right of decimal point.

| Parameter | u   | g   | r   | i   | z   | redshift |
|-----------|-----|-----|-----|-----|-----|----------|
| u         | 1.0 | 0.85| 0.69| 0.6 | 0.55| 0.16     |
| g         | 0.85| 1.0 | 0.96| 0.91| 0.88| 0.41     |
| r         | 0.69| 0.96| 1.0 | 0.98| 0.97| 0.44     |
| i         | 0.6 | 0.91| 0.98| 1.0 | 0.98| 0.43     |
| z         | 0.55| 0.88| 0.97| 0.98| 1.0 | 0.42     |
| redshift  | 0.16| 0.41| 0.44| 0.43| 0.42| 1.0      |

Statistical properties of the dataset parameters are given at the table 2. It’s clearly seen that nature of the redshift parameter differs from all other values. Its diapason also differs. Parameters u, g, r, i, z are of the same kind. They have got close values and close statistical measures: minima and maxima values, standard deviation and mean values (the table 2).

| Parameter | Amount of non-zero values | Mean  | Standard deviation | Minimum | Maximum |
|-----------|---------------------------|-------|--------------------|---------|---------|
| u         | 10000                     | 18.62 | 0.83               | 12.99   | 19.60   |
| g         | 10000                     | 17.37 | 0.95               | 12.80   | 19.92   |
| r         | 10000                     | 16.84 | 1.07               | 12.43   | 24.80   |
| i         | 10000                     | 16.58 | 1.14               | 11.95   | 28.18   |
| z         | 10000                     | 16.42 | 1.20               | 11.61   | 22.83   |
| redshift  | 10000                     | 0.14  | 0.39               | ≤0.01   | 5.36    |

The redshift values have got different statistical properties and diapason that’s way all parameters have been scaled. New value has got the form shown in the expression (1):
Here $x$ is an old value, $x'$ is the transformed one, $\mu$ is its mean value and $\sigma$ is its standard deviation. After this step standard deviation of all the values is 1, mean value is 0. Minima and maxima are shown in the table 3.

### Table 3. Statistical measures of parameters in the Sloan Digital Sky Survey DR14 dataset after normalization.

| Parameter | Minimum | Maximum |
|-----------|---------|---------|
| $u$       | -6.80   | 1.18    |
| $g$       | -4.84   | 2.69    |
| $r$       | -4.13   | 7.46    |
| $i$       | -4.06   | 10.16   |
| $z$       | -4.00   | 5.33    |
| redshift  | -0.38   | 13.40   |

All parameters have got close values now. It’s possible to use them inside of one classifier.

In the binary classification problem (the classifier function $a(x)$ can return only 1 and -1 for each object $x$) constructed classifier can return correct values and it can make mistakes. Their types are presented in the table 4.

### Table 4. Types of classifiers’ correct responses and errors in the binary classification task (classifiers’ responses $a(x)$ are situated in the rows and correct labels $y$ in the columns).

| $a(x)$  | $y = 1$          | $y = -1$          |
|---------|------------------|-------------------|
| $a(x) = 1$ | True positive (TP) | False positive (FP) |
| $a(x) = -1$ | False negative (FN) | True negative (TN) |

The measures of classifiers’ quality usually are precision (2) and recall (3), also $F_\beta, \beta = 1$ value (4) is implemented:

$$\text{precision} = \frac{TP}{TP + FP}, \quad (2)$$

$$\text{recall} = \frac{TP}{TP + FN}, \quad (3)$$

$$F_\beta = (1 + \beta^2) \frac{\text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}. \quad (4)$$

The AUC_PR measure denotes area under the precision-recall curve [18].

### 3. Experiments

The described dataset has been processed with a few classifiers. Their results are compared in the last part.

#### 3.1. Decision tree

One of the simplest classifier in the data science is decision tree or CART [19, 20]. At the same time this sort of trees has got one delicious advantage: the structure of such tree might be explained. Researcher can tell which parameters or their combinations are important in this classification task. New object processing is just a consequence of questions about its characteristics. Decision trees can reach
large height. To avoid overfitting and to understand the structure of the classifier one can leave only some higher levels of the tree (pruning) [19].

The diapason bounds dividing classes are shown in the table 5 (conditions in the nodes of three highest levels are presented in the column Parameter Bounds). The Gini coefficient for appropriate node of the decision tree is presented in the last column.

Table 5. Bounds of classes (star, galaxy or quasar).

| Type of object | Parameter Bounds | Gini coefficient value |
|----------------|------------------|------------------------|
| Galaxy         | redshift ∈ [−0.36, 0.23], \( g \leq 1.65 \) | 0.03 \( i \leq \) −0.06. |
| Quasar         | redshift > 0.23, | 0.06 \( i \leq \) −0.06. |
| Star           | redshift < −0.36. | 0.01 \( i \leq \) −0.06. |

After the training procedures had been completed the test set has been handled with the built classifier. Its precision metrics are shown in the table 6. The main diagonal contains correct. At the cell with row index \( i \) and column index \( j \) there’s the ratio of the responses when the classifier have chosen class \( i \) but the correct response has been class \( j \).

Table 6. Mean confusion matrix of the CART decision tree classifier.

| Correct class label | Galaxy | 0.992 | 0.006 | 0.002 |
|---------------------|--------|-------|-------|-------|
| Quasar              | 0.076  | 0.924 | 0     | 0     |
| Star                | 0      | 0     | 1     | 0     |
| Galaxy              | Quasar | Star  |       |       |

The metrics of the decision tree classifier for Sloan Digital Sky Survey dataset are shown in the table 7. The last row contains mean value of each metrics for all classes.

Table 7. Quality of the CART decision tree classifier.

| Classes and metrics | Precision | Recall | \( F_1 \) | AUC-PR |
|---------------------|-----------|--------|-----------|--------|
| Galaxy              | 0.986     | 0.993  | 0.990     | 0.983  |
| Quasar              | 0.973     | 0.921  | 0.946     | 0.903  |
| Star                | 0.998     | 1.0    | 0.999     | 0.998  |
| Mean                | 0.986     | 0.971  | 0.978     | 0.961  |

In the table 6 one can see that the galaxies and stars at an acceptable level. At the same time 8% of quasars are identified as galaxies. Applications of this algorithms to star classification are available in [8, 9]. The random forest algorithm results aren’t shown because this classifier has got almost the same accuracy. Implementation of the random forest to star classification is discussed in [12]. Improvement of classification algorithms with use of special genetic algorithm is presented in [16] and is going to be tested in future work.

3.2. Logistic regression

The main idea of this algorithm is to construct a hyperplane or a set of hyperplanes dividing objects of each class. As well some kind of regularization is also included in the optimization task of the logistic regression classifier. Thus, the coefficients in of the hyperplane are supposed not to be very large. That’s
one of the ways to struggle against overfitting. The result of the logistic regression classifier [19, 21] is shown in the tables 8 and 9.

Table 8. Mean confusion matrix of the logistic regression classifier.

| Correct class label | Galaxy | 0.990 | 0.003 | 0.007 |
|---------------------|--------|-------|-------|-------|
| Quasar              | 0.051  | 0.949 |      0 |       |
| Star                | 0      | 0     | 1     |       |

Table 9. Quality of the logistic regression classifier.

| Classes and metrics | Precision | Recall | \(F_1\) | AUC-PR |
|---------------------|-----------|--------|---------|--------|
| Galaxy              | 0.991     | 0.987  | 0.989   | 0.984  |
| Quasar              | 0.985     | 0.949  | 0.967   | 0.940  |
| Star                | 0.988     | 1.0    | 0.994   | 0.988  |
| Mean                | 0.988     | 0.979  | 0.983   | 0.971  |

The metrics of the logistic regression classifier for Sloan Digital Sky Survey dataset are shown in the table 9. The last row contains mean value of each metrics for every class.

One can conclude that logistic regression divides galaxies and stars well. In case of quasars 5% of objects are classified as galaxies.

3.3. Naïve Bayes classifier
The result of the naïve Bayes classifier [19, 21] is shown in the tables 10 and 11.

Table 10. Mean confusion matrix of the naïve Bayes classifier.

| Correct class label | Galaxy | 0.982 | 0.014 | 0.004 |
|---------------------|--------|-------|-------|-------|
| Quasar              | 0.059  | 0.941 |      0 |       |
| Star                | 0.002  | 0.007 | 0.991 |       |

Table 11. Quality of the naïve Bayes classifier.

| Classes and metrics | Precision | Recall | \(F_1\) | AUC-PR |
|---------------------|-----------|--------|---------|--------|
| Galaxy              | 0.988     | 0.983  | 0.985   | 0.979  |
| Quasar              | 0.898     | 0.941  | 0.919   | 0.850  |
| Star                | 0.996     | 0.992  | 0.994   | 0.991  |
| Mean                | 0.960     | 0.972  | 0.966   | 0.940  |

The metrics of naïve Bayes classifier for Sloan Digital Sky Survey dataset are shown in the table 11. The last row contains mean value of each metrics for every class.

One can conclude that the naïve Bayes classifier divides galaxies and stars well like other classifiers shown above. Though its measures are a bit worse. The majority of errors are also made while processing quasars.
3.4. Enhanced dataset procession

In the second part of the experiment additional columns have been added to the dataset: all the parameters squared, products of all pairs of parameters and logarithms of each one. After these functions had been counted all the values have been normalized according to (1). There are a lot of values close to 0 in the redshift column. So, 1 was added to its value before application of logarithm.

Classifiers mentioned above have been used to handle this enhanced dataset. At the same time it’s clear that stars are classified very good almost with each algorithm. Majority of errors have occurred while choosing between galaxies and stars (all the stars have been identified with the logistic regression and the decision tree algorithms). Because of that all the stars have been deleted from the dataset. The binary classification (quasar or galaxy) task has been solved.

Results of the decision tree are shown in the table 12.

**Table 12. Quality of the CART decision tree classifier at the enhanced dataset.**

| Correct class label | Galaxy | 0.99 | 0.008 |
|---------------------|--------|------|-------|
| Quasar              | 0.082  | 0.94 |
|                     | Galaxy | Quasar |

Predicted class label: $F_1$ value of this classifier if 0.9608. Height of the best decision tree is 7.

Results of the random forest classifier are shown in the table 13.

**Table 13. Quality of the random forest tree classifier at the enhanced dataset.**

| Correct class label | Galaxy | 0.99 | 0.008 |
|---------------------|--------|------|-------|
| Quasar              | 0.079  | 0.94 |
|                     | Galaxy | Quasar |

Predicted class label: $F_1$ value of this classifier if 0.9742. Height of the best random forest ensemble is 9. Its results are a bit better than the decision tree ones. But improvement isn’t dramatical.

The logistic regression classifier has made better differentiation. It’s presented in the table 14.

**Table 14. Quality of the logistic regression classifier at the enhanced dataset.**

| Correct class label | Galaxy | 1    | 0.002 |
|---------------------|--------|------|-------|
| Quasar              | 0.048  | 0.95 |
|                     | Galaxy | Quasar |

Predicted class label: $F_1$ value of this classifier if 0.982. The quality of the classifier has achieved 98%.

The gradient boosting is an improved algorithm using boosting methods. It uses wrongly classified objects at the first stage as a core of the train set at next stages of learning [19 – 21]. Usually it has got better results than simple methods demonstrated above. $F_1$ value of the gradient boosting classifier here is 0.9687. This result confirms that the data itself isn’t differentiable. To build good classifier in this task one has to use kurtosis values that aren’t presented in the handled dataset.

**Table 15. Quality of the gradient boosting classifier at the enhanced dataset.**

| Correct class label | Galaxy | 0.99 | 0.006 |
|---------------------|--------|------|-------|
| Quasar              | 0.064  | 0.93 |
|                     | Galaxy | Quasar |

Predicted class label
3.5. **Principal component analysis**

Principal component analysis with special kernel functions and without them has been applied to the enhanced dataset. There hasn’t been any quality improvement. But it has become clear that the presented data isn’t enough to solve the task. From the physical point of view researchers need kurtosis values available in other datasets [22]. Values of frequency allow to differentiate light from galaxies with wide spectrum and narrow peaks which come from quasars. Plots of the first principal components are presented below at the figures 1 – 3 to show these features of the dataset. To save space only three principal components are used (89% of explained variance). The first plot denotes connection between the first and the second components, the second one shows mutual behaviour of the first and the third components. The last one is the plot of the second and the third principal component. To explain 95% of variance in the enhanced dataset one has to use five components.

**Figure 1.** Plot of dependence between the first and second principal components of the enhanced dataset (quasars are marked with black color, galaxies are gray points)

**Figure 2.** Plot of dependence between the first and third principal components of the enhanced dataset (quasars are marked with black color, galaxies are gray points)

**Figure 3.** Plot of dependence between the second and third principal components of the enhanced dataset (quasars are marked with black color, galaxies are gray points)

4. **Conclusion**

The Sloan Digital Sky Survey DR14 dataset has been investigated in this paper. It includes statistical information about visual characteristics of stars, galaxies and quasars. The classifiers based on different learning algorithms have been built. In the paper it has been shown that the simplest classifiers are enough to solve this task. Though there are a lot of papers describing neural networks implementation to solve such problems these methods look too complicated for star classification.

A set of simple learning models have been used to classify this dataset. The main problem is to distinguish galaxies and quasars. The stars have been identified almost at 100% level with the majority of classifiers. Mean values of all classification characteristics are greater than 94%. So, in the first part of the experiment it’s unnecessary to construct the random forest classifier or other because it almost doesn’t improve these results.

In the second part the classification of galaxies and quasars has been under investigation. The logistic regression classifier has achieved the best results (98%). At the same time one can conclude that the 100% values are impossible in this task. To illustrate this hypothesis the plots of the first principal components have been presented. From the physical point of view one needs kurtosis values that aren’t included in the handled dataset. The statistical values of the light (form of kurtosis) are enough to distinguish quasars. If the necessary data is added into this set it would be possible to solve the task at the level of 100% precision.

The most of the classifiers behave like black boxes. The decision tree structure can explain the solution. The borders of each class in terms of the dataset parameters are presented in the table 5 and can be interpreted physically.
References

[1] Finch A and Said J L 2018 Galactic rotation dynamics in f(T) gravity EPJ C Particles and Fields 78 560 doi: 10.1140/epjc/s10052-018-6028-1

[2] Sadovnikova E V and Shatina A V 2018 Evolution of the rotational motion of a satellite with flexible viscoelastic rods on the elliptic orbit Russian Technological Journal 6(4) 89-104 url: https://rtj.mirea.ru/upload/media/library/4dc/RTZH_4_2018_89_104.pdf

[3] Abbott B P et al. 2017 Gravitational Waves and Gamma-Rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A The Astrophysical J Let 848 (2) L13 doi: 10.3847/2041-8213/aa920c

[4] Lee K J et al. 2013 PEACE: pulsar evaluation algorithm for candidate extraction – a software package for post-analysis processing of pulsar survey candidates Monthly Notices of the Royal Astronomical Soc. 433 688–94 doi:10.1093/mnras/stt758

[5] Wang Y-C and Li M-T and Pan Z-C and Zheng J-H 2019 Pulsar candidate classification with deep convolutional neural networks Research in Astronomy and Astrophysics 19(9) 133 doi: 10.1088/1674–4527/19/9/133

[6] Wang L and Jin J and Jiang Y and Shen Y A 2019 Method for Weak Pulsar Signal Detection Combining the Bispectrum and a Deep Convolutional Neural Network The Astrophysical J 873 17 doi:10.3847/1538-4357/ab0308

[7] Zhu W W et al. 2014 Searching for pulsars using image pattern recognition The Astrophysical J, 781 117 (12pp) doi:10.1088/0004-637X/781/2/117

[8] Vasconcellos E C, et al. 2011 Decision tree classifiers for star/galaxy separation The Astrophysical J 141(6) 189 doi:10.1088/0004-6256/141/6/189

[9] Ball N M, Brunner R J and Myers A D 2006 Robust machine learning applied to astronomical data sets. Star-galaxy classification of the Sloan digital sky survey DR3 using decision trees The Astrophysical J 650 497-509

[10] Ackermann M et al. 2012 A statistical approach to recognizing source classes for unassociated sources in the first Fermi-Lat catalog The Astrophysical Journal 753(1) 83 doi:10.1088/0004-637X/753/1/83

[11] Saz Parkinson P M, Xu H, Yu P L H, Salvetti D, Marelli M and Falcone A D 2016 Classification and ranking of Fermi-Lat gamma-ray sources from the 3FGL catalog using machine learning techniques The Astrophysical J 820(1) 8 doi:10.3847/0004-637X/820/1/8

[12] Farrell S A, Murphy T and Lo K K 2015 Autoclassification of the variable 3xmm sources using the random forest machine learning algorithm The Astrophysical J 813 28 doi:10.1088/0004-637X/813/1/28

[13] Weaver W B 2000 Spectral classification of unresolved binary stars with artificial neural networks The Astrophysical J 541 298-305

[14] Richards G T et al. 2004 Efficient photometric selection of quasars from the Sloan digital sky survey: 100,000 z < 3 quasars from data release one The Astrophysical J Supplement Series 155 257-69

[15] Richards G T et al. 2015 Bayesian high-redshift quasar classification from optical and mid-ir photometry The Astrophysical J Supplement Series 219(2) 39 doi:10.1088/0067-0049/219/2/39

[16] Anfyorov M A 2019 Genetic clustering algorithm Russian Technological Journal 7(6) pp134-50 https://doi.org/10.32362/2500-316X-2019-7-6-134-150

[17] Sloan Digital Sky Survey DR14 Classification of Stars, Galaxies and Quasars. Retrieved from: https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey

[18] Davis J and Goardrich M 2006 The Relationship Between Precision-Recall and ROC Curves Proc. of the 23rd Int. Conf. on Machine Learning (Pittsburgh, PA, USA)

[19] James G, Witten D, Hastie T and Tibshirani R 2015 An introduction to statistical learning with applications in R (Springer-Verlag New York) p 426 doi: 10.1007/978-1-4614-7138-7

[20] Breiman L, Freidman J H, Olshen R A and Stone C J 1984 Classification And Regression Trees
[21] Hastie T, Tibshirani R and Friedman J 2009 *The elements of statistical learning* (Springer-Verlag New York) p 533

[22] Predicting a pulsar star Retrieved from: https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star