

Математическое моделирование

Mathematical modeling

УДК 004.94

<https://doi.org/10.32362/2500-316X-2021-9-3-66-77>

НАУЧНАЯ СТАТЬЯ

О выборе метода распознавания астрономических объектов на основе анализа исходных данных, полученных по программе Sloan Digital Sky Survey DR14

В.А. Голов,
Д.А. Петрусевич [®]

МИРЭА – Российский технологический университет, Москва, 119454 Россия

[®] Автор для переписки, e-mail: petrusevich@mirea.ru, petrdenis@mail.ru

Резюме. В работе проведен анализ набора данных Sloan Digital Sky Survey DR14, в котором в несколько этапов измерений собраны статистические данные о различных астрономических объектах. На поверхности Земли расположено много телескопов, собирающих данные об объектах в небе. На околоземной орбите и в космосе (обычно в точках Лагранжа систем Земля – Луна, Солнце – Земля) расположены или запланированы к размещению телескопы, следящие за небом в разных диапазонах. Значительный объем данных приводит к необходимости статистической обработки этого потока информации, а также к построению автоматических классификаторов по типу объекта. В работе представлены результаты предварительной обработки данных и работы различных видов классификаторов в задаче определения типа астрономического объекта из набора данных Sloan Digital Sky Survey DR14 (звезда, квазар или галактика) на основе нескольких распространенных метрик. Рассмотрены алгоритмы дерева принятий решений, логистическая регрессия, наивный «байесовский» классификатор и ансамбли классификаторов. Показано, что классификация подобных наборов данных может быть проведена без привлечения сложных систем машинного обучения (таких, как нейронные сети). Сделаны выводы об особенностях применения алгоритмов машинного обучения к этой задаче. В некоторых случаях работа классификаторов может быть интерпретирована с точки зрения физики. Точность построенных в работе классификаторов (согласно метрикам, учитывающим несбалансированность классов) достигает 90% и может считаться удовлетворительной и для того, чтобы считать задачу решенной, и для того, чтобы использовать структуру классификаторов для объяснения результатов классификации с точки зрения физики.

Ключевые слова: Sloan Digital Sky Survey DR14, анализ данных, машинное обучение, дерево принятия решений, логистическая регрессия

• Поступила: 13.04.2020 • Доработана: 29.08.2020 • Принята к опубликованию: 09.03.2021

Для цитирования: Голов В.А., Петрусевич Д.А. О выборе метода распознавания астрономических объектов на основе анализа исходных данных, полученных по программе Sloan Digital Sky Survey DR14. *Российский технологический журнал*. 2021;9(3):66–77. <https://doi.org/10.32362/2500-316X-2021-9-3-66-77>

Прозрачность финансовой деятельности: Никто из авторов не имеет финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

RESEARCH ARTICLE

Data analysis methods in astronomic objects classification (Sloan Digital Sky Survey DR14)

Vladislav A. Golov,
Denis A. Petrushevich @

MIREA – Russian Technological University, Moscow, 119454 Russia
@ Corresponding author, e-mail: petrushevich@mirea.ru, petrdenis@mail.ru

Abstract. In the paper Sloan Digital Sky Survey DR14 dataset was investigated. It contains statistical information about many astronomical objects. The information was obtained within the framework of the Sloan Digital Sky Survey project. There are telescopes at the Earth surface, at the Earth orbit and in the Lagrange points of some systems (Earth–Moon, Sun–Earth). The telescopes gain information in different frequency ranges. The large quantity of statistical information leads to the demand for analytical algorithms and systems capable of making classification. Such information is marked up well enough to build machine learning classification systems. The paper presents the results of a number of classifiers. The handled data contains measures of three types of astronomical objects of the Sloan Digital Sky Survey DR14 dataset (star, quasar, galaxy). The CART decision tree, logistic regression, naïve Bayes classifiers and ensembles of classifiers (random forest, gradient boosting) were implemented. Conclusions about special features of each machine learning classifier trained to solve this task are made at the end of the paper. In some cases, classifiers' structure can be explained physically. The accuracy of the classifiers built in this research is more than 90% (metrics F1, precision and recall are implemented, because the classes are unbalanced). Taking these values into account classification task is supposed to be successfully solved. At the same time, the structure of classifiers and importance of features can be used as a physical explanation of the solution.

Keywords: Sloan Digital Sky Survey DR14, data analysis, machine learning, decision tree, logistic regression

• Submitted: 13.04.2020 • Revised: 29.08.2020 • Accepted: 09.03.2021

For citation: Golov V.A., Petrushevich D.A. Data analysis methods in astronomic objects classification (Sloan Digital Sky Survey DR14). *Rossiiskii tekhnologicheskii zhurnal = Russian Technological Journal*. 2021;9(3):66–77 (in Russ.). <https://doi.org/10.32362/2500-316X-2021-9-3-66-77>

Financial disclosure: No author has a financial or property interest in any material or method mentioned.

The authors declare no conflicts of interest.

ВВЕДЕНИЕ

Анализ данных и машинное обучение – это новые области знаний, которые оказали влияние на решение задач во многих сферах знаний. Долгое время исследование астрономических объектов на основе фиксаций изменения блеска объектов и их перемещения на последовательных снимках, а также на основе анализа спектра, проводилось вручную. Но сейчас объем накопленных данных и скорость их получения такова, что самостоятельная обработка этого потока информации занимает слишком много времени и усилий. Более того, от исследователя это и не требуется. Вопрос поиска объектов, обладающих заданными свойствами, перешел в область обработки большого набора статистических данных.

На основе накопленных данных решается широкий круг задач, например:

- уточнение динамики объектов (система планета – спутник; гравитационная структура галактик; описание динамики объектов вокруг сверхмассивной черной дыры Стрелец А*, центра галактики Млечный Путь) [1, 2];
- классификация звезд (поиск пульсаров и отделение шума от сигнала пульсара) [3–6];
- поиск экзопланет (чаще всего основываясь на кратковременном периодическом изменении блеска звезды);
- в последнее время существенное внимание уделяется вопросам анализа гравитационных волн и сведению результатов наблюдений за гравитационным полем и электромагнитных волн [7].

С точки зрения информационных технологий, в основном, в области обработки астрономических данных встречаются две разновидности задач. Накопленный статистический материал без использования графической информации исследуется методами математической статистики и анализа данных [4, 8–12]. Наборы данных, основанные на множестве снимков неба или определенного объекта, анализируются методами компьютерной графики, спектрального анализа и машинного зрения [5, 6, 13–15].

Представленное исследование посвящено анализу на примере набора данных Sloan Digital Sky Survey DR14 [16]. В рамках работы показано, что представленный набор данных может быть обработан без использования классификаторов, требующих сложного процесса обучения и настройки (таких, как нейронные сети).

В работах [4, 5] для решения задачи классификации строятся нейронные сети глубокого обучения (deep convolutional neural networks). Их структура и методы обучения сложны, требуют опыта и искусства настройки. Авторы [4] отмечают преобладание количества зафиксированных звезд и галактик над числом квазаров (как в наборе данных, так и в целом). По этой причине требуется либо использовать специальные метрики качества классификации (как сделано в представленном исследовании), либо дополнять класс квазаров искусственными примерами, сгенерированными с помощью специально обученных нейронных сетей. В исследовании [5] авторы анализируют частотные характеристики сигнала с помощью нейронных сетей. Сигнал пульсара характеризуется узким диапазоном вокруг частоты, соответствующей максимуму отношения сигнал/шум. Использование сетей глубокого обучения направлено на распознавание типа сигнала на основе частотных характеристик.

В исследовании [6] построена двухуровневая комбинация различных классификаторов. На первом этапе конструируются комбинации «простых» классификаторов (метод опорных векторов) и классификаторов на основе сверточных нейронных сетей (convolutional neural networks, CNN). На втором этапе классификатор, получающий их ответы, обучается так, чтобы максимизировать метрику качества F_1 . Используемая схема похожа на идею построения «сильных» классификаторов на основе множества «слабых» (беггинг) [17, 18].

В работе [8] построено решающее дерево для анализа одной из предыдущих версий исследуемого в представленной статье набора данных Sloan Digital Sky Survey DR7 [19]. Авторы не проводят сравнение различных алгоритмов классификации, а используют один из самых простых – решающее дерево [17, 18]. При этом проведено сравнение различных

алгоритмов построения решающего дерева на основе метрик качества классификации, анализируются настраиваемые на этапе обучения классификатора параметры (такие, как высота дерева), вводятся новые параметры, например, комбинация исходных параметров набора данных. Представленная работа посвящена сравнению различных алгоритмов классификации в приложении к астрономическим объектам [16], в том числе исследуется классификатор на основе решающего дерева.

В исследованиях [10, 11] проанализирован каталог астрономических объектов, опубликованный американским агентством NASA [20]. Проведена многоклассовая классификация на основе информации о различных источниках излучения в виде частотных характеристик. В рамках этих исследований для классификации использовано решающее дерево (так же, как в работе [8]), также построены классификаторы на основе логистической регрессии, на основе бустинга произведено усиление классификаторов, использующих логистическую регрессию [17, 18].

Идея усиления простых классификаторов (обладающих точностью выше 50%) заключается в настройке их комбинации таким образом, чтобы повысить точность классификации (рассчитываемой на основе обработки обучающей выборки объектов). Концепции беггинга и бустинга – наиболее известные алгоритмы построения таких комбинаций [17, 18]. В работе [12] для классификации астрономических объектов используется классификатор на основе «случайного леса»: с помощью беггинга комбинируется работа множества решающих деревьев.

На основе представленного обзора можно сделать следующие выводы:

- в большинстве проанализированных работ сравниваются два алгоритма построения классификаторов или за основу берется классификатор определенного вида;
- классификаторы на основе выбранных алгоритмов настраиваются вручную либо используются концепции усиления простых классификаторов на основе беггинга или бустинга с целью повышения качества классификации;
- в обзоре не выявлено исследований, в которых сравнивается более двух классификаторов, построенных на основе разных алгоритмов.

Актуальность исследования состоит в том что, построение классификаторов астрономических объектов на основе нейронных сетей (в том числе, глубокого обучения) является излишне сложным путем решения задачи классификации, т.к. требует процедуры обучения и настройки конфигурации и параметров сети (которая зачастую является искусством исследователей). Простые классификаторы могут дать точность, сопоставимую с результатами,

полученными с помощью систем на основе нейронных сетей. Кроме того, в статье приведено сравнение точностей классификаторов, построенных авторами, а также с классификаторами, построенными в исследованиях, вошедших в литературный обзор.

Цель работы – сравнение классификаторов, построенных на основе различных алгоритмов, не затрагивающих обучение нейронных сетей.

1. НАБОР ДАННЫХ И МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

Информация в наборе данных Sloan Digital Sky Survey DR14 [16] представляет собой 10 000 наблюдений космоса. Каждое наблюдение описывается 17 столбцами признаков и 1 столбцом класса, который идентифицирует его как представителя одного из классов: звезду (4152 объектов), галактику (4998 объектов) или квазар (850 объектов). Так как большая часть столбцов является описанием номера объекта или координатами, описывающими положение объекта в пространстве, в рамках данной работы интерес представляют 6 признаков столбцов и один столбец класса.

Объекты описываются набором из 6 признаков:

1. Фотометрическая система u, g, r, i, z – широкополосная пятицветная фотометрическая система. Подразумевается значение звездной величины в каждом фильтре.
2. Красное смещение $redshift$ – сдвиг излучения в сторону красной части спектра (длинных волн).

Корреляция между признаками представлена в таблице 1. Диагональные элементы, которые отвечают за корреляцию параметров с собой, равны 1 и не указаны. Все результаты в таблице 1 и далее в выражениях в тексте работы приведены с точностью до 0.01, за исключением метрик качества классификаторов (приведены до 3-го знака после запятой).

Таблица 1. Корреляция между основными признаками в наборе данных Sloan Digital Sky Survey DR14

Признак	u	g	r	i	z	$redshift$
u		0.85	0.69	0.6	0.55	0.16
g	0.85		0.96	0.91	0.88	0.41
r	0.69	0.96		0.98	0.97	0.44
i	0.6	0.91	0.98		0.98	0.43
z	0.55	0.88	0.97	0.98		0.42
$redshift$	0.16	0.41	0.44	0.43	0.42	

Статистические характеристики признаков набора данных представлены в таблице 2. Природа параметра $redshift$ отличается от остальных признаков, диапазон его значений отличается существенно.

Признаки u, g, r, i, z явно имеют схожую природу и близкие значения минимумов, максимумов, стандартного отклонения и среднего значения.

Таблица 2. Статистические характеристики признаков в наборе данных Sloan Digital Sky Survey DR14

Признак	Количество ненулевых записей	Среднее значение	Стандартное отклонение	Минимум	Максимум
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

В связи с тем, что значения признака $redshift$ выделены и отличаются по диапазону значений, проведено нормирование признаков методом масштабирования:

$$x' = \frac{x - \mu}{\sigma}, \quad (1)$$

где x – значение признака до преобразования; x' – значение параметра после преобразования; μ – его среднее значение; σ – среднеквадратическое отклонение.

Среднее значение всех параметров после нормирования приобрело значение 0, стандартное отклонение равно 1. Значения минимумов и максимумов приведены в таблице 3.

Таблица 3. Статистические характеристики признаков в наборе данных Sloan Digital Sky Survey DR14 после нормирования

Признак	Минимум	Максимум
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

Таким образом, все признаки приобрели значения из одинакового диапазона, что делает возможным их совместное использование в классификаторах.

В задаче бинарной классификации (принадлежность двум классам: $\{1, -1\}$) возможны как

правильные ответы $a(x)$ классификатора, так и ошибки, типы которых представлены в таблице 4.

Таблица 4. Типы правильных ответов и ошибок классификатора в задаче бинарной классификации (в строках ответы классификатора $a(x)$, в столбцах – верные метки y классов)

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

Оценка точности как соотношения между правильными ответами классификатора и общим числом протестированных примеров (accuracy) зависит от соотношения объектов разных классов и не может быть использована, если классы не сбалансированы по числу примеров [17, 18]:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (2)$$

Для оценки качества классификатора введены понятия точности (precision) и полноты (recall) [17, 18], не зависящие от соотношения количества объектов в каждом классе:

$$precision = \frac{TP}{TP + FP}, \quad (3)$$

$$recall = \frac{TP}{TP + FN}. \quad (4)$$

На основе этих оценок вводится мера F :

$$F_\beta = (1 + \beta^2) \frac{precision \times recall}{\beta^2 precision + recall}. \quad (5)$$

При $\beta = 1$ в формуле (5) получаем среднее гармоническое с коэффициентом 2.

Также используется площадь под кривой AUC-PR на графике точность – полнота (precision – recall) [21].

2. ЭКСПЕРИМЕНТЫ

В экспериментальной части исследования на представленном наборе данных было протестировано несколько алгоритмов классификации. Их результаты сравниваются в пункте 2.6.

2.1. Дерево решений

Одним из самых простых классификаторов в области анализа данных является дерево (принятия) решений [17, 18, 22]. Зачастую за его простотой скрывается и положительная

характеристика – возможность проследить, какие параметры (или их комбинации) оказались важны при построении классификатора, а сама классификация фактически превращается в ответы на набор вопросов (каждый из которых можно представить в узле бинарного дерева). В связи с тем, что дерево может иметь очень большую высоту для сложных задач, в целях борьбы с переобучением (overfitting), а также, чтобы понять смысл построенного классификатора, деревья решений усекают по высоте (pruning) [17].

Границы диапазонов параметров, соответствующие каждому из трех классов, полученные с помощью усеченного дерева решений, представлены в таблице 5.

Таблица 5. Границы признаков для определения класса (звезда, квазар или галактика)

Тип объекта	Ограничения на значения признаков
Галактика	$redshift \in [-0.36, 0.23]$, $g \leq 1.65$
Квазар	$redshift > 0.23$, $i \leq -0.06$
Звезда	$redshift < -0.36$

Коэффициент Джини [17] для классов: 1) 0.03 – для галактик; 2) 0.06 – для квазаров; 3) 0.01 – для звезд.

После обучения классификатора его качество было проверено на тестовой выборке. Точность классификации представлена в таблице 6. На главной диагонали расположена доля правильных ответов классификатора. В ячейке с номером строки i , номером столбца j стоит доля ответов классификатора в такие моменты, когда правильный класс имеет индекс j , а классификатор дал ответ в виде класса i .

Таблица 6. Матрица ошибок (mean confusion matrix) классификатора, построенного на основе дерева решений

Правильная метка класса	Галактика	0.992	0.006	0.002
	Квазар	0.076	0.924	0
	Звезда	0	0	1
		Галактика	Квазар	Звезда
		Предсказанная метка класса		

В таблице 7 приведены результаты работы классификатора, согласно различным метрикам, с точностью до 0.001 [21, 22].

Таблица 7. Метрики качества классификатора на основе дерева решений

Классы и метрики	Precision (точность)	Recall (полнота)	F_1	AUC_PR
Галактика	0.986	0.993	0.990	0.983
Квазар	0.973	0.921	0.946	0.903
Звезда	0.998	1.0	0.999	0.998
Среднее	0.986	0.971	0.978	0.961

Последняя строка содержит оценку среднего значения метрики качества по всем классам.

По результатам работы классификатора видно, что решающее дерево распознает верно все галактики и звезды, но порядка 8% квазаров классифицирует как галактики.

2.2. Логистическая регрессия

Суть этого классификатора заключается в построении гиперплоскостей в пространстве признаков, разделяющих области каждого класса. В задачу оптимизации, решаемую при обучении классификатора, обычно включают регуляризацию весов (коэффициентов гиперплоскостей). Фактически, это метод борьбы с переобучением (overfitting). Результаты классификатора, основывающегося на логистической регрессии [17, 18], представлены в таблицах 8, 9.

Таблица 8. Матрица ошибок (mean confusion matrix) классификатора, построенного на основе логистической регрессии

Правильная метка класса	Галактика	0.990	0.003	0.007
	Квазар	0.051	0.949	0
	Звезда	0	0	1
		Галактика	Квазар	Звезда
		Предсказанная метка класса		

В таблице 9 приведены результаты работы классификатора, согласно различным метрикам, с точностью до 0.001.

Таблица 9. Метрики качества классификатора на основе логистической регрессии

Классы и метрики	Precision (точность)	Recall (полнота)	F_1	AUC_PR
Галактика	0.991	0.987	0.989	0.984
Квазар	0.985	0.949	0.967	0.940
Звезда	0.988	1.0	0.994	0.988
Среднее	0.988	0.979	0.983	0.971

По результатам работы классификатора видно, что с помощью логистической регрессии удается правильно распознать все галактики и звезды, но порядка 5% квазаров было отнесено к классу галактик.

2.3. «Наивный» Байесовский классификатор

В работе этого классификатора делается предположение о том, что между признаками нет взаимозависимости. Хотя такое условие выполняется крайне редко, этот метод остается хорошим инструментом для исследования данных. Используя теорему Байеса для вероятности сопоставления объекта с параметрами $X = (x_1, \dots, x_n)$ и метки класса y_i , получим:

$$P(y_i | x_1, \dots, x_n) = \frac{P(y_i)P(x_1, \dots, x_n | y_i)}{P(x_1, \dots, x_n)}. \quad (6)$$

Здесь $P(x_1, \dots, x_n | y_i)$ – вероятность встретить объект $X = (x_1, \dots, x_n)$ среди объектов класса y_i , оцениваемая при обучении [17, 18].

На этапе классификации необходимо выбрать класс с максимальной оценкой вероятности $\arg \max(P(y_i | x_1, \dots, x_n))$. Следует отметить, что знаменатель при классификации не меняется, поэтому достаточно оценить только числитель дроби (6). В связи с тем, что существует несколько меток классов y_i , получаем:

$$P(x_1, \dots, x_n | y) = \prod_{i=1}^{|Y|} P(x_1, \dots, x_n | y_i). \quad (7)$$

Здесь $|Y|$ – мощность множества меток при классификации. Зачастую оцениваемое выражение логарифмируют и получают сумму (8). Требуется отобрать класс, для которого значение этого выражения максимально.

$$Y = \arg \max \left(\log(P(y_i)) + \sum_{i=1}^{|Y|} \log(P(x_1, \dots, x_n | y_i)) \right). \quad (8)$$

Результаты классификатора, основывающегося на логистической регрессии [18], представлены в таблицах 10, 11.

Таблица 10. Матрица ошибок (mean confusion matrix) «наивного» Байесовского классификатора

Правильная метка класса	Галактика	0.982	0.014	0.004
	Квазар	0.059	0.941	0
	Звезда	0.002	0.007	0.991
		Галактика	Квазар	Звезда
		Предсказанная метка класса		

В таблице 11 приведены результаты работы классификатора, согласно различным метрикам, с точностью до 0.001.

Таблица 11. Метрики качества «наивного» Байесовского классификатора

Классы и метрики	Precision (точность)	Recall (полнота)	F_1	AUC_PR
Галактика	0.988	0.983	0.985	0.979
Квazar	0.898	0.941	0.919	0.850
Звезда	0.996	0.992	0.994	0.991
Среднее	0.960	0.972	0.966	0.940

Классификатор работает для звезд и галактик с высоким качеством (оценка 93% чуть ниже, чем у логистической регрессии), но при этом большинство ошибок снова допущено при классификации квазаров.

2.4. Добавление новых параметров в набор данных

Точность классификаторов, построенных в пунктах 3.1–3.3, высока, но ошибки возникают в большинстве случаев при разделении галактик и квазаров. Более того, классификаторы на основе решающего дерева и логистической регрессии выделяют все звезды в наборе данных. По этой причине произведен переход к решению задачи бинарной классификации: в наборе данных оставлены только примеры галактик и квазаров, которые надо разделить. При обработке данных применен метод главных компонент.

На рис. 1–3 представлены зависимости трех главных компонент (описывают 99% разброса данных) друг от друга (квазары отмечены черным, галактики – серым).

Как видно из рис. 1–3, примеры галактик и квазаров смешаны в пространстве признаков, поэтому были добавлены дополнительные параметры:

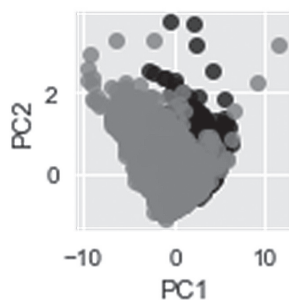


Рис. 1. Зависимость первой и второй главных компонент в наборе данных

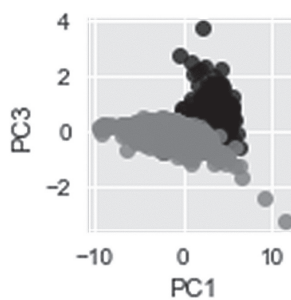


Рис. 2. Зависимость первой и третьей главных компонент в наборе данных

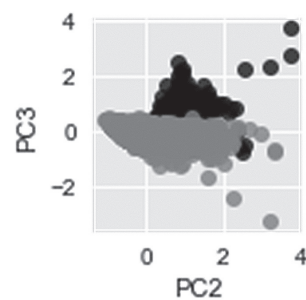


Рис. 3. Зависимость второй и третьей главных компонент в наборе данных

квадраты всех признаков (u , g , r , i , z , $redshift$), все возможные попарные произведения исходных признаков, а также их логарифмы. Функции применялись к исходным значениям параметров, затем проводилась нормализация согласно (1). Значения некоторых параметров перед нормализацией были близки к 0 ($redshift$), поэтому предварительно к ним добавлялось значение 1.

К расширенному набору данных было применено несколько классификаторов.

Решающее дерево с наилучшей мерой $F_1 = 0.9608$ имеет высоту 7. Качество классификации представлено в таблице 12.

Таблица 12. Матрица ошибок решающего дерева (CART) на расширенном наборе данных

Правильная метка класса	Галактика	0.99	0.008
	Квazar	0.082	0.94
		Галактика	Квazar
		Предсказанная метка класса	

Случайный лес (англ. термин *random forest*) включает ансамбль классификаторов на основе решающих деревьев с усреднением результата классификации [17, 18]. Наилучший результат показал классификатор с ограничением по высоте деревьев – 9 (таблица 13).

Таблица 13. Матрица ошибок случайного леса на расширенном наборе данных

Правильная метка класса	Галактика	0.99	0.008
	Квazar	0.079	0.94
		Галактика	Квazar
		Предсказанная метка класса	

Значение меры F_1 случайного леса равно 0.9742. Его результаты чуть лучше, чем у решающего дерева, но существенного улучшения нет.

Классификатор на основе логистической регрессии позволил получить лучшую точность классификации. Его результаты представлены в таблице 14. Значение меры F_1 достигло 0.982.

Таблица 14. Матрица ошибок классификатора на основе логистической регрессии на расширенном наборе данных

Правильная метка класса	Галактика	1	0.002
	Квazar	0.048	0.95
		Галактика	Квazar
		Предсказанная метка класса	

Классификатор на основе градиентного бустинга [17, 18, 22] обычно имеет более высокое качество, отличаясь при этом и сложностью, затратным обучением с вычислительной точки зрения. Основная идея бустинга – использование на некотором этапе обучения классификатора объектов, неправильно классифицированных на предыдущих шагах. Значение меры F_1 при использовании градиентного бустинга на этих данных равно 0.9687. Существенного улучшения достичь не удалось. Результаты представлены в таблице 15.

Таблица 15. Матрица ошибок классификатора на основе градиентного бустинга на расширенном наборе данных

Правильная метка класса	Галактика	0.99	0.006
	Квazar	0.064	0.93
		Галактика	Квazar
		Предсказанная метка класса	

2.5. Применение метода главных компонент

Применение метода главных компонент показало, что для описания 96% разброса в данных необходимо 5 компонент. Попарные графики их зависимостей друг от друга приведены на рис. 4–13 (квазары отмечены черным, галактики – серым).

Как видно, галактики и квазары снова оказались смешаны, поэтому построить классификатор с точностью 100% по такому набору данных не удастся. Следует отметить, что, согласно физике вопроса, квазары имеют высокий коэффициент эксцесса – куртозис (kurtosis). Именно этот момент и отделяет их излучение от света, испущенного другими объектами. В наборе данных [23] как раз приведены подобные значения для квазаров, поэтому можно констатировать, что в исследуемом наборе данных просто нет необходимых для классификации значений, а есть косвенные характеристики.

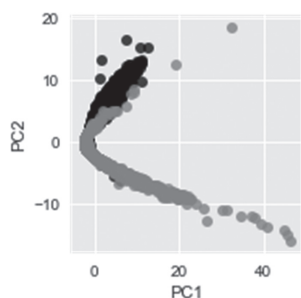


Рис. 4. Зависимость первой и второй главных компонент в расширенном наборе данных

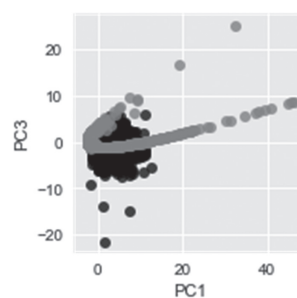


Рис. 5. Зависимость первой и третьей главных компонент в расширенном наборе данных

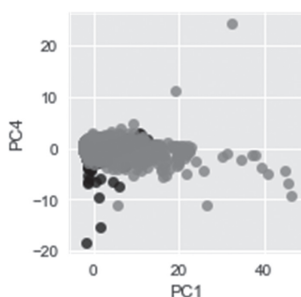


Рис. 6. Зависимость первой и четвертой главных компонент в расширенном наборе данных

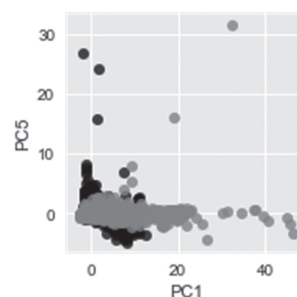


Рис. 7. Зависимость первой и пятой главных компонент в расширенном наборе данных

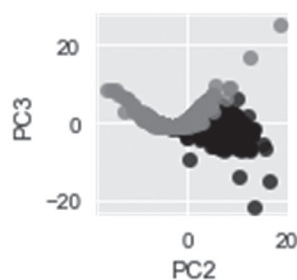


Рис. 8. Зависимость второй и третьей главных компонент в расширенном наборе данных

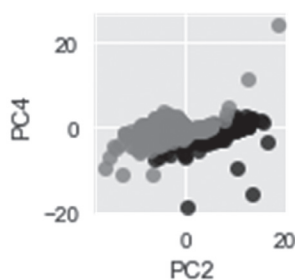


Рис. 9. Зависимость второй и четвертой главных компонент в расширенном наборе данных

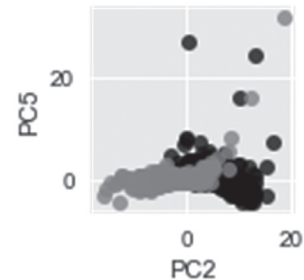


Рис. 10. Зависимость второй и пятой главных компонент в расширенном наборе данных

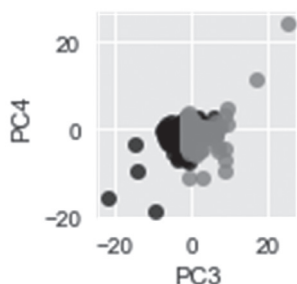


Рис. 11. Зависимость третьей и четвертой главных компонент в расширенном наборе данных

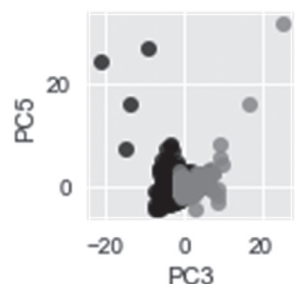


Рис. 12. Зависимость третьей и пятой главных компонент в расширенном наборе данных

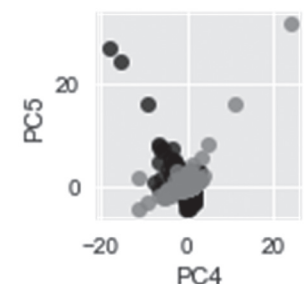


Рис. 13. Зависимость четвертой и пятой главных компонент в расширенном наборе данных

2.6. Сравнение точности классификаторов

Характеристики качества классификаторов, построенных в исследованиях [4, 5, 8, 11, 12, 14, 15], приведены в таблице 16.

В данной работе на исходном наборе данных максимальные значения характеристик качества

классификации были получены с использованием логистической регрессии ($F_1 = 0.983$, см. таблицу 9) и дерева решений ($F_1 = 0.978$, см. таблицу 7). На расширенном наборе данных лучшие результаты были получены при работе с классификаторами на основе случайного леса ($F_1 = 0.974$, см. таблицу 13) и логистической регрессии ($F_1 = 0.982$, см. таблицу 14), что не уступает результатам, приведенным в таблице 16.

Таблица 16. Точность построенных классификаторов в исследованиях, упоминающихся в литературном обзоре

Наименование литературного источника	Анализируемый в работе набор данных	Оценка качества классификатора
Wang Y.-C., Li M.-T., Pan Z.-C., Zheng J.-H. Pulsar candidate classification with deep convolutional neural networks [4]	Информация о трех пульсарах и искусственно сгенерированные примеры	$accuracy > 0.9$
Wang L., Jin J., Jiang Y., Shen Y. A Method for weak pulsar signal detection combining the bispectrum and a deep convolutional neural network [5]	Классификация пульсаров на основе набора данных Predicting a pulsar star [23]	$F_1 = 0.96$
Vasconcellos E.C., de Carvalho R.R., Gal R.R., LaBarbera F.L., Capelato H.V., Frago Campos Velho H., Trevisan M., Ruiz R.S.R. Decision tree classifiers for star/galaxy separation [8]	Классификация квазаров на основе набора данных Digital Sky Survey DR7 [19]	Полнота (recall) в трех разных диапазонах звездных величин: >99%, 68 – 85%, 42 – 74%

Окончание табл. 16

Наименование литературного источника	Анализируемый в работе набор данных	Оценка качества классификатора
Saz Parkinson P.M., Xu H., Yu P.L.H., Salvetti D., Marelli M., Falcone A.D. Classification and ranking of Fermi-Lat gamma-ray sources from the 3FGL catalog using machine learning techniques [11]	Многоклассовая классификация источников (пульсары, квазары, сверхновые и туманности, двойные системы и другие объекты из каталога Fermi-LAT 3FGL Catalog) [20]	$accuracy = 0.96$
Farrell S.A., Murphy T., Lo K.K. Autoclassification of the variable 3xmm sources using the random forest machine learning algorithm [12]	Многоклассовая классификация источников (двойные системы, пульсары и другие объекты из каталога 3XMM) [24]	$accuracy = 0.96$
Richards G.T., <i>et al.</i> Efficient photometric selection of quasars from the Sloan digital sky survey: 100,000 $z < 3$ quasars from data release one [14]	Классификация квазаров на основе набора данных Digital Sky Survey DR1 [25]	$accuracy = 0.95$
Richards G.T., <i>et al.</i> Bayesian high-redshift quasar classification from optical and mid-ir photometry [15]	Классификация квазаров на основе набора данных Digital Sky Survey DR3 [26] и графической информации набора данных AllWise [27]	$accuracy = 0.93$

ЗАКЛЮЧЕНИЕ

В представленном исследовании проведена обработка набора данных Sloan Digital Sky Survey DR14 (kaggle.com) по классификации астрономических объектов на квазары, галактики и звезды. Построено несколько простых классификаторов, основанных на решающих деревьях, логистической регрессии, а также «наивный» байесовский классификатор. На следующем этапе были применены ансамбли классификаторов: случайный лес, градиентный бустинг. В ходе литературного обзора выявлено множество работ, в которых аналогичные задачи решаются на основе нейронных сетей различных конфигураций и сложности [4, 5, 13]. Результаты, полученные в этом исследовании, показывают, что для решения подобной задачи достаточно использовать наиболее простые алгоритмы машинного обучения. Точность построенных в исследовании классификаторов сравнивалась с характеристиками классификаторов, построенных в работах [4, 5, 8, 11, 12, 14, 15]. Максимальная точность достигнута при настройке классификатора на основе логистической регрессии – $F_1 = 98\%$, что не уступает результатам, полученным в исследованных работах.

Задача классификации была решена с помощью нескольких простых моделей. Основной проблемой стало разделение галактик и квазаров. Большинство алгоритмов распознало практически все звезды в тренировочных наборах верно. Решающее дерево и логистическая регрессия показали результат в 100%

при определении звезд. Средние значения метрик классификации превосходят 94%. По этой причине было решено не строить классификатор на основе случайного леса.

Далее галактики и квазары были выделены в отдельный набор данных, и решалась задача бинарной классификации. Наилучший результат показала логистическая регрессия – $F_1 = 98\%$. При обработке данных стало ясно, что построить классификатор с точностью 100% в этой задаче не получится. С одной стороны, применение метода главных компонент показало (рис. 1–3), что объекты перемешаны в пространстве признаков. Введение вторых степеней параметров, их попарных произведений, логарифмов, применение метода главных компонент (с ядром и без) не изменило эту ситуацию (рис. 4–13). С точки зрения физики, свет квазаров можно выделить на основе анализа частотных характеристик (высокий эксцесс, куртозис). Исследуемый набор данных [16] (в отличие от [21]) не содержит этих величин. При добавлении в набор данных соответствующей информации построение классификаторов с точностью 100% представляется возможным.

С методической точки зрения представленная работа может служить учебным пособием при построении простых классификаторов и часто применяемых на практике ансамблей классификаторов.

Вклад авторов. Все авторы в равной степени внесли свой вклад в исследовательскую работу.

Authors' contribution. All authors equally contributed to the research work.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Finch A., Said J.L. Galactic rotation dynamics in $f(T)$ gravity. *Eur. Phys. J. C.* 2018;78:560. <https://doi.org/10.1140/epjc/s10052-018-6028-1>
2. Садовникова Е.В., Шатина А.В. Эволюция вращательного движения спутника с гибкими вязкоупругими стержнями на эллиптической орбите. *Российский технологический журнал*. 2018;6(4):89–104. <https://doi.org/10.32362/2500-316X-2018-6-4-89-104> [Sadovnikova E.V., Shatina A.V. Evolution of the rotational motion of a satellite with flexible viscoelastic rods on the elliptic orbit. *Rossiiskii Tekhnologicheskii Zhurnal = Russian Technological Journal*. 2018;6(4):89–104 (in Russ.).]
3. Lee K.J., Stovall K., Jenet F.A., *et al.* PEACE: pulsar evaluation algorithm for candidate extraction – a software package for post-analysis processing of pulsar survey candidates. *Mon. Not. R. Astron. Soc.* 2013;433(1):688–694. <https://doi.org/10.1093/mnras/stt758>
4. Wang Y.-C., Li M.-T., Pan Z.-C., Zheng J.-H. Pulsar candidate classification with deep convolutional neural networks. *Res. Astron. Astrophys.* 2019;19(9):133. <https://doi.org/10.1088/1674-4527/19/9/133>
5. Wang L., Jin J., Jiang Y., Shen Y. A Method for weak pulsar signal detection combining the bispectrum and a deep convolutional neural network. *Astrophys. J.* 2019;873(1):17. <https://doi.org/10.3847/1538-4357/ab0308>
6. Zhu W.W., Berndsten A., Madsen E.C., *et al.* Searching for pulsars using image pattern recognition. *Astrophys. J.* 2014;781(2):117. <https://doi.org/10.1088/0004-637X/781/2/117>
7. Abbott B.P., Abbot R., Abbott T.D., *et al.* Gravitational waves and gamma-rays from a binary neutron star merger: GW170817 and GRB 170817A. *Astrophys. J. Lett.* 2017;848(2):L13. <https://doi.org/10.3847/2041-8213/aa920c>
8. Vasconcellos E.C., de Carvalho R.R., Gal R.R., LaBarbera F.L., Capelato H.V., Frago Campos Velho H., Trevisan M., Ruiz R.S.R. Decision tree classifiers for star/galaxy separation. *Astrophys. J.* 2011;141(6):189. <https://doi.org/10.1088/0004-6256/141/6/189>
9. Ball N.M., Brunner R.J., Myers A.D. Robust machine learning applied to astronomical data sets. I. Star-galaxy classification of the Sloan digital sky survey DR3 using decision trees. *Astrophys. J.* 2006;650(1):497–509. <https://doi.org/10.1086/507440>
10. Ackermann M., Ajello M., Allafort A., *et al.* A statistical approach to recognizing source classes for unassociated sources in the first Fermi-LAT catalog. *Astrophys. J.* 2012;753(1):83. <https://doi.org/10.1088/0004-637X/753/1/83>
11. Saz Parkinson P.M., Xu H., Yu P.L. H., Salvetti D., Marelli M., Falcone A.D. Classification and ranking of Fermi-LAT gamma-ray sources from the 3FGL catalog using machine learning techniques. *Astrophys. J.* 2016;820(1):8. <https://doi.org/10.3847/0004-637X/820/1/8>
12. Farrell S.A., Murphy T., Lo K.K. Autoclassification of the variable 3XMM sources using the random forest machine learning algorithm. *Astrophys. J.* 2015;813(1):28. <https://doi.org/10.1088/0004-637X/813/1/28>
13. Weaver W.B. Spectral classification of unresolved binary stars with artificial neural networks. *Astrophys. J.* 2000;541(1):298–305. <https://doi.org/10.1086/309425>
14. Richards G.T., Nichol R.C., Gray A., *et al.* Efficient photometric selection of quasars from the Sloan digital sky survey: 100,000 $z < 3$ quasars from data release one. *Astrophys. J. Suppl. Ser.* 2004;155(2):257–269. <https://doi.org/10.1086/425356>
15. Richards G.T., Myers A.D., Peters C.M., *et al.* Bayesian high-redshift quasar classification from optical and mid-ir photometry. *Astrophys. J. Suppl. Ser.* 2015;219(2):39. <https://doi.org/10.1088/0067-0049/219/2/39>
16. Sloan Digital Sky Survey DR14. Classification of stars, galaxies and quasars. Available from URL: <https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey>
17. James G., Witten D., Hastie T., Tibshirani R. *An introduction to statistical learning with applications in R*. New York: Springer-Verlag; 2015. 426 p. <https://doi.org/10.1007/978-1-4614-7138-7>
18. Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning*. 2nd ed. New York: Springer-Verlag; 2009. 763 p. ISBN 978-0-387-84857-0
19. Sloan digital sky survey data release 7. Available from URL: <http://classic.sdss.org/dr7>
20. Fermi-LAT 3FGL Catalog. Available from URL: https://fermi.gsfc.nasa.gov/ssc/data/access/lat/4yr_catalog/3FGL-table/#aitoff
21. Davis J., Goadrich M. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, PA, USA; 2006. p. 233–240. <https://doi.org/10.1145/1143844.1143874>
22. Breiman L., Friedman J.H., Olshen R.A., Stone C.J. *Classification and regression trees*. Monterey, CA, USA: Wadsworth & Brooks/Cole Advanced Books & Software; 1984. 358 p. ISBN: 0534980548
23. Predicting a pulsar star. Available from URL: <https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star>
24. VizieR online data catalog: XMM-Newton Serendipitous Source Catalogue 3XMM-DR8 (XMM-SSC, 2018). Available from URL: <https://ui.adsabs.harvard.edu/abs/2019yCat.9055....0R/abstract>
25. Sloan digital sky survey data release 1. Available from URL: <http://classic.sdss.org/dr1/>
26. Sloan digital sky survey data release 3. Available from URL: <http://classic.sdss.org/dr3/>
27. The wide-field infrared survey explorer at IPAC. The AllWISE Data Release. Available from URL: <http://wise2.ipac.caltech.edu/docs/release/allwise/>

Об авторах

Голов Владислав Александрович, студент группы КМБ-01-16 кафедры высшей математики Института кибернетики ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: golov.v.a@yandex.ru.

Петрусевич Денис Андреевич, к.ф.-м.н., доцент кафедры высшей математики Института кибернетики ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: petrusevich@mirea.ru, petrdenis@mail.ru. ResearcherID: AAA-6661-2020, Scopus Author ID: 55900513600, <https://orcid.org/0000-0001-5325-6198>

About the authors

Vladislav A. Golov, Student, Higher Mathematics Department, Institute of Cybernetics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: golov.v.a@yandex.ru.

Denis A. Petrusevich, Cand. Sci. (Phys.–Math.), Associate Professor, Higher Mathematics Department, Institute of Cybernetics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: petrusevich@mirea.ru, petrdenis@mail.ru. ResearcherID: AAA-6661-2020, Scopus Author ID: 55900513600, <https://orcid.org/0000-0001-5325-6198>