Факультет космических исследований Московского государственного университета имени М. В. Ломоносова

Государственный астрономический институт им. П. К. Штернберга Московского государственного университета имени М.В. Ломоносова

МГУ имени М.В. Ломоносова

Извлечение признаков из кривых блеска астрономических источников

А. Д. Лаврухина[1], К. Л. Маланчев[2,3]

[1]Факультет космических исследований, Московский государственный университет им. М.В. Ломоносова, Ленинские горы, 2-й учебный корпус, Москва, 119991, Россия
[2]Государственный астрономический институт им. П.К. Штернберга, Университетский пр-т. 13, Москва, 119234, Россия
[3]Департамент Астрономии, Иллинойсский Университет в Урбане—Шампейне, Урбана, США

Абстракт

Современные астрономические обзоры содержат информацию о сотнях миллионах кривых блеска переменных астрономических источников, например, релиз данных Zwicky Transient Facility Data Release 3 (ZTF DR3) содержит миллиарды кривых блеска. При решении задач классификации [1] или поиска аномалий [3] в таких больших объемах данных используются методы машинного обучения. Обычно кривые блеска не используются напрямую, вместо чего каждый источник представляется набором признаков, которые наилучшим образом описывают свойства его переменности. Часть из этих признаков может быть взята из работ по ручной разметке некоторых классов объектов. Такие признаки хорошо описывают физические свойства переменности отдельных типов объектов. Другие признаки используются для описания свойств кривой блеска как неравномерного временного ряда. В данной работе мы представляем новую библиотеку на языках Python и Rust, предназначенную для извлечения признаков из кривых блеска переменных астрономических источников.

Тестирование библиотеки

В качестве тестовых данных использовались 1855 кривых блеска цефеид и 2394 кривых блеска карликовых новых в красной фотометрической полосе. Кривые блеска взяты из релиза данных ZTF DR3. На Рис. 1 изображены графики кривых блеска для представителей обоих типов.

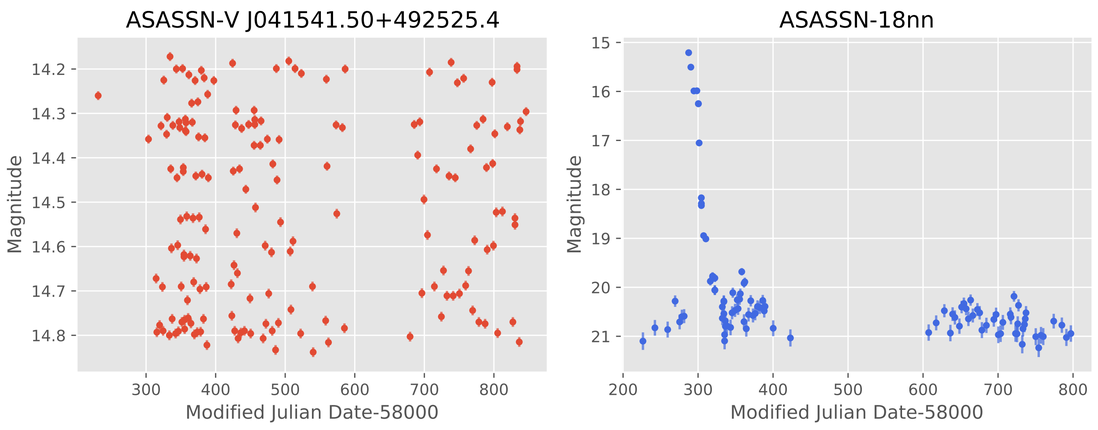

Рис. 1: Графики кривых блеска для цефеиды ASASSN-V J041541.50+492525.4 (красным) и карликовой новой ASASSN-18nn (синим)

Пусть кривая блеска задана набором из N наблюдений $\{t_i, m_i, \delta_i\}$, где t_i — время наблюдения, m_i — наблюдаемая звездная величина, δ_i — ошибка наблюдения звездной величины.

Тестовая задача заключалась в классификации кривых блеска (каждой кривой блеска в результате работы алгоритма присваивается значение соответствующего класса, т.е. тип звезды). Для этого из каждой кривой блеска были извлечены 25 признаков. Одним из методов решения задач классификации является алгоритм случайного леса [2].
Для решения тестовой задачи классификации были построены 3 модели, в каждой из которой случайный лес содержал 100 деревьев с максимальной глубиной 5.

• lavrukhina.ad@gmail.com
• https://github.com/light-curve

Каждая модель обучалась на одинаковом наборе данных, состоящем из 85% исходных данных, и тестировалась на оставшихся 15% данных. Перед разбиением на тренировочную и тестовую выборку данные были перемешаны. Все модели были построены при помощи пакета SCIKIT-LEARN 0.24.1. [4].

1. В первой модели на вход алгоритма случайного леса подавались все 25 признаков.
2. Для второй модели была проведена процедура уменьшения размерности пространства признаков при помощи метода главных компонент (principal component analysis, PCA) до 15 компонент. Количество компонент было выбрано исходя из процента объясненной дисперсии модели (15 компонент объясняет 99% дисперсии).
3. Для третьей модели был произведен отбор 15 признаков при помощи sequential feature selection.

Для оценки качества моделей использовались метрики precision (точность), recall (полнота), ROC AUC (площадь под ROC-кривой) и accuracy (доля правильных ответов). Качество модели тем лучше, чем больше значение соответствующей метрики.

	Metric			
	precision	recall	ROC AUC	accuracy
model 1	0.955	0.958	0.990	0.958
model 2	0.961	0.936	0.985	0.950
model 3	0.962	0.968	0.992	0.966

Таблица 1: Сравнительная таблица метрик для описанных моделей

Как видно из Табл. 1 все модели имеют примерно одинаковое качество классификации тестовых данных. Уменьшение размерности пространства признаков до 15 не ухудшает качество модели, выбор признаков при помощи sequential feature selection позволяет добиться прироста качества.

Выводы

Использование рассмотренных признаков, вероятно, позволит успешно применять алгоритмы машинного обучения для решения задач классификации или поиска аномалий. В дальнейшем мы планируем реализовать функции для извлечения других признаков, а также разработать новые признаки, подходящие для задач классификации. Кроме того, в планах валидация признаков при помощи решения тестовых задач классификации с применением машинного обучения.

Ссылки

[1] D.-W. Kim, P. Protopapas, C. A. L. Bailer-Jones, Y.-I. Byun, S.-W. Chang, J.-B. Marquette, and M.-S. Shin. The epoch project. *Astronomy & Astrophysics*, 566:A43, Jun 2014. ISSN 1432-0746. doi: 10.1051/0004-6361/201323252. URL `http://dx.doi.org/10.1051/0004-6361/201323252`.

[2] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.

[3] K. L. Malanchev, M. V. Pruzhinskaya, V. S. Korolev, P. D. Aleo, M. V. Kornilov, E. E. O. Ishida, V. V. Krushinsky, F. Mondon, S. Sreejith, A. A. Volnova, A. A. Belinski, A. V. Dodin, A. M. Tatarnikov, and S. G. Zheltoukhov. Anomaly detection in the Zwicky Transient Facility DR3. *arXiv e-prints*, art. arXiv:2012.01419, Dec. 2020.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-02-00779. Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета "Фундаментальные и прикладные исследования космоса"