DOI: 10.15514/ISPRAS-2019-31(2)-3

Гибридная модель для эффективного обнаружения аномалий в кратковременных последовательностях кривых блеска GWAC и аналогичных наборах данных

И. Сан, ORCID: 0000-0003-0545-3175 <sunying1304@126.com>
З. Жао, ORCID: 0000-0002-3638-0290 <xiaoemma6@163.com>
С. Ма, ORCID: 0000-0001-6622-6318 <matnt2008@126.com>
Чж. Ду, ORCID: 0000-0002-8435-1611 <duzh@tsinghua.edu.cn>
Факультет компьютерных наук и технологий, Университет Цинхуа, Китай

Аннотация. Раннее оповещение во время обзора неба дает важную возможность обнаруживать одиночные планеты с малой массой. В статье представлен гибридный метод, в котором комбинируется модель ARIMA (интегрированная модель авторегрессии – скользящего среднего), рекуррентные нейронные сети (RNN) LSTM (нейронная сеть с блоками долго-кратковременной памяти) и GRU (управляемый рекуррентный нейрон), обеспечивающий возможность поиска кратковременных событий микролинзирования (ML) в режиме реального времени на основе данных, получаемых путем высокочастотной широкоугольной съемки звездного неба. Метод обеспечивает мониторинг всех наблюдаемых кривых блеска и выявление событий ML на ранних стадиях. Экспериментальные результаты показывают, что гибридные модели обеспечивают бо́льшую точность и требуют меньше времени на настройку параметров. ARIMA + LSTM и ARIMA + GRU могут повысить точность на 14,5% и 13,2% соответственно. При обнаружении аномалий в кривых блеск, GRU может достичь почти того же результата, что и LSTM, затрачивая на 8% меньшее время. Те же модели применимы и к набору данных ЭКГ в базах данных MIT-BIH по аритмии с похожим паттерном аномалий, и в обоих случаях мы можем сократить на 40% временя, которое требуется исследователям для настройки модели, с сохранением 90% точности.

Ключевые слова: гравитационное линзирование; рекурентные нейронные сети; ARIMA; предупреждения и прогнозы на основе временных рядов

Для цитирования. Сан И., Жао З., Ма С., Ду Чж. Гибридная модель для эффективного обнаружения аномалий в кратковременных последовательностях кривых блеска GWAC и аналогичных наборах данных. Труды ИСП РАН, том 31, вып. 2, 2019 г., стр. 33-40. DOI: 10.15514/ISPRAS-2019-31(2)-3

Благодарности. Исследование частично поддерживалось Программой базовых исследований и разработок КНР (грант No.2016YFB1000602), Базовой лабораторией космической астрономии и технологии Национальной астрономической обсерватории Китайской академии наук, Национальным фондом естественных наук КНР (гранты 61440057, 61272087, 61363019, 61073008, 11690023) и Фондом исследовательского центра МОЕ в области дистанционного образования (грант No. 2016ZD302)

Hybrid Model for Efficient Anomaly Detection in Short-timescale GWAC Light Curves and Similar Datasets

Y. Sun, ORCID: 0000-0003-0545-3175 <sunying1304@126.com>
Z. Zhao, ORCID: 0000-0002-3638-0290 <xiaoemma6@163.com>
X. Ma, ORCID: 0000-0001-6622-6318 <matnt2008@126.com>
Z. Du, ORCID: 0000-0002-8435-1611 <duzh@tsinghua.edu.cn>
Department of Computer Science and Technology, Tsinghua University

Abstract. Early warning during sky survey provides a crucial opportunity to detect low-mass, free-floating planets. In particular, to search short-timescale microlensing (ML) events from high-cadence and wide- field survey in real time, a hybrid method which combines ARIMA (Autoregressive Integrated Moving Average) with LSTM (Long-Short Time Memory) and GRU (Gated Recurrent Unit) recurrent neural networks (RNN) is presented to monitor all observed light curves and identify ML events at their early stages. Experimental results show that the hybrid models perform better in accuracy and less time consuming of adjusting parameters. ARIMA+LSTM and ARIMA+GRU can achieve improvement in accuracy by 14.5% and 13.2%, respectively. In the case of abnormal detection of light curves, GRU can achieve almost the same result as LSTM with less time by 8%. Same models are also applied to MIT-BIH Arrhythmia Databases ECG dataset with similar abnormal pattern and we yield from both sets that we can reduce up to 40% of time consuming for researchers to adjust the model to 90% accuracy.

Keywords: gravitational lensing; recurrent neural networks; ARIMA; time series prediction and alarming

For citation: Sun Y., Zhao Z., Ma X., Du Z. Hybrid Model for Efficient Anomaly Detection in Short-timescale GWAC Light Curves and Similar Datasets. Trudy ISP RAN/Proc. ISP RAS, vol. 31, issue 1, 2019. pp. 33-40 (in Russian). DOI: 10.15514/ISPRAS-2019-31(2)-3

Acknowledgements. This research is supported in part by Key Research and Development Program of China (No.2016YFB1000602), the Key Laboratory of Space Astronomy and Technology, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 100012, China, National Natural Science Foundation of China (Nos. 61440057,61272087,61363019 and 61073008,11690023), MOE research center for online education foundation (No 2016ZD302).

1. Введение

Астрономия является отправной точкой взрыва информации, и это первая область, которая встретилась с проблемой больших данных [1]. В этой статье мы используем данные мини-GWAC (Ground-based Wide Angle Camera, наземная широкоугольная камера) в качестве образцов для обучения и тестирования. Мы исследуем проблему поиска в реальном времени гравитационных ML-событий кратковременного масштаба в огромном наборе кривых блеска, применяя гибридные модели ARIMA-LSTM и ARIMA-GRU. Мы также пытаемся выработать подход, который мог бы применяться в области финансов и других областях, подобно исследованиям, выполняемым в Корнелле [2]. Эксперименты для оценки производительности этих двух моделей выполнялись на наборе данных мини-GWAC.

Мы также применяем свои модели к базе данных аритмии MIT-BIH, характеристики которой схожи с набором данных GWAC: аномальные ситуации похожи одна на другую, а нормальные ситуации ведут себя беспорядочно. База данных аритмии MIT-BIH [5] содержит 48 получасовых выдержек из двухканальных амбулаторных записей ЭКГ, полученных от 47 субъектов, которые обследовались в лаборатории аритмии BIH между 1975 и 1979 годами.

2. Гибридные модели ARIMA-LSTM и ARIMA-GRU

Данные детектирования света, полученные с помощью мини-GWAC, содержат более 900 файлов временных рядов для каждой планеты. Очевидно, что стоит предположить наличие у этих данных линейной, так и нелинейной частей [3]. Таким образом, мы можем представить данные следующим образом:

$$x_t = L_t + N_t + \varepsilon_t$$

Здесь L_t представляет линейность данных в момент времени t, а N_t обозначает нелинейность. Значение ε_t представляет погрешность. В предыдущей работе [4] на линейных задачах отличные результаты показала интегрированная модель авторегрессии – скользящего среднего (Autoregressive Integrated Moving Average, ARIMA), обеспечивая в целом точность более 85%. Это традиционный метод прогнозирования временных рядов. С другой стороны, модель долго-кратковременной памяти (LSTM) может обнаруживать в наборе данных нелинейные тренды. Наша гибридная модель ARIMA-LSTM позволяет обнаруживать как линейные, так и нелинейные тренды.

3. Методика проведения экспериментов

3.1 Наборы данных и среда проведения экспериментов

Наборы данных GWAC и mini GWAC: Данные GWAC до недавнего времени не были открыты, поэтому наши алгоритмы тестируются на наборе данных mini-GWAC. Система Mini-GWAC состоит из 12 комплектов широкоугольных камер. Она была построена и размещена в обсерватории Синлун Национальной астрономической обсерватории. В этой статье мы используем данные мини-GWAC в качестве примеров для обучения и тестирования. Для каждой планеты набор данных мини-GWAC содержит в среднем 980 текстовых файлов. В каждом файле мы можем получить 900 данных, составляющих часть временных рядов.

База данных аритмии MIT-BIH– набор данных ЭКГ: Мы обнаружили в наборе данных ЭКГ характеристики, схожие с набором данных GWAC: аномальные ситуации похожи одна на другую, а нормальные ситуации ведут себя беспорядочно. База данных аритмии MIT-BIH [5] содержит 48 получасовых выдержек из двухканальных амбулаторных записей ЭКГ, полученных от 47 субъектов, которые обследовались в лаборатории аритмии BIH между 1975 и 1979 годами.

3.2 Алгоритм

Алгоритм 1 представляет собой краткое описание нашего метода. На первом этапе для прогнозирования приблизительных результатов используется ARIMA. Размер окна составляет 20% от длины массива. Для каждого окна производится результат. После получения результата окно перемещается на шаг вперед, чтобы предсказать следующее значение. Однако между реальностью и предсказанием имеются некоторые остатки. Поэтому для более точного прогнозирования остатки рассчитываются и используются в качестве входных данных на втором этапе, где они являются обучающими наборами в RNN. Таким образом, на втором шаге обучающие наборы используются для прогнозирования нелинейной части. Наконец, на последнем этапе окончательные прогнозы представляют собой сумму значений, предсказанных ARIMA, и остатков. Такие прогнозы более точны, чем прогнозы, сделанные только с помощью ARIMA.

```
1.   Result = []
2.   Resid = emptylist
3.   for dataset in datasets do
4.     Predict = []
5.     Models = emptylist
6.     Order = arima.aicminorder
7.     if unstable: then
8.        Model = diff(model)
9.     end if
10.    Model = fitarima(dataset, order)
11.    Add residual[0] to Resid
12.    Add predict[0] to Predict
13.    Add predict to Result
14.  end for
15.  Save Result[-1], Resid
```

Алгоритм 1. Гибридная модель ARIMA-LSTM/ARIMA-GRU
Algorithm 1. ARIMA-LSTM/ARIMA-GRU Hybrid Model

4. Результаты экспериментов и их оценка

4.1 Результаты тестирования на наборе данных mini-GWAC

В этом подразделе мы выявляем три аспекта эффективности модели: точность, затрачиваемое время и сложность вычислений. Точность определяется тем, насколько рано выдаются оповещения и насколько часто появляются ложные предсказания. Другими словами, модель должна быть одновременно точной при прогнозировании и чувствительной к аномальным случаям. На рис. 1 точка оповещения помечена вертикальной красной линией. Чем меньше время оповещения, тем лучше работает модель. Ложное предсказание – это расхождение предсказанного и реального значений.

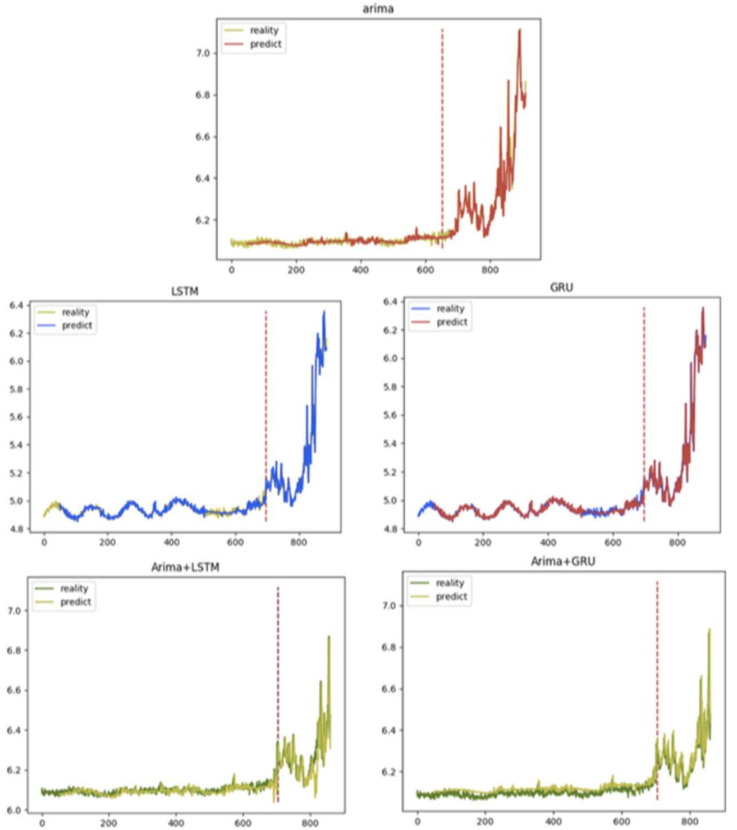

Рис. 1. Тестирование разных алгоритмов на наборе данных mini-GWAC
Fig. 1. Testing different algorithms on mini-GWAC dataset

Результаты для набора данных мини-GWAC приведены в табл. 1.

a) При кратковременном прогнозирования событий гравитационного микролинзирования с использованием набора данных мини-GWAC GRU обучается быстрее, чем LSTM.

b) Своевременность оповещения у проверенных методов различается мало. Точность немного лучше у гибридных моделей. LSTM ведет себя более надежно, чем GRU.

c) Методы GRU проще и, следовательно, их легче модифицировать, добавляя, например, вентили при потребности ввода в сеть дополнительных данных. Это приводит к сокращению времени обучения и сложности вычислений.

d) ARIMA может достигать меньшего времени оповещения и времени работы, но имеется высокая частота ложных предсказаний. За счет сокращения времени работы на 15% гибридные модели ARIMA и LSTM или GRU могут улучшить точность на 14,5% и 13,2% соответственно.

Табл. 1. Результаты оценки разных алгоритмов на наборе данных mini-GWAC
Table 1. Evaluation results of different algorithms on the mini-GWAC dataset

Алгоритмы	Точность / время оповещения	Время выполнения
ARIMA	81.60% / 41.7%	0.349 сек.
LSTM	93.72% / 42.6%	0.478 сек.
GRU	93.28% / 43.3%	0.440 сек.
ARIMA-LSTM	96.11% / 42.2%	0.406 сек.
ARIMA-GRU	94.83% / 42.8%	9.413 сек.

4.2 Результаты тестирования на базе данных аритмии MIT-BIH

Результаты всех моделей показаны на рис. 2, и хотя результаты не так убедительны, как в предыдущем подразделе, они позволяют прийти к некоторым выводам.

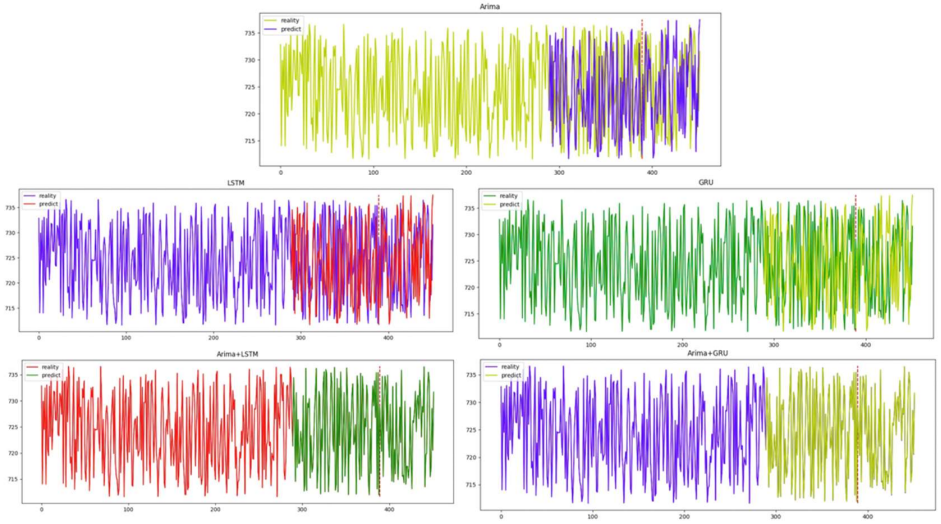

Рис. 1. Тестирование разных алгоритмов на наборе данных ECG
Fig. 2. Test on ECG dataset of different algorithms

a) Для набора данных ECG ARIMA-LSTM работает лучше, чем ARIMA-GRU, обеспечивая большую точности, и меньшее время выполнения.

b) Из-за параметров RNN точность LSTM и GRU составляет всего около 50%, но в сочетании с ARIMA они могут обеспечить точность свыше 90%.

c) Время выполнения намного больше, чем для наборов данных GWAC.

Табл. 1. Результаты оценки разных алгоритмов на наборе данных ECG
Table 1. Evaluation results of different algorithms on the ECG dataset

Алгоритмы	Точность / время оповещения	Время выполнения
ARIMA	91.47% / 37.9%	4.12 сек.
LSTM	52.22% / 30.8%	7.61 сек.
GRU	50.30% / 30.8%	7.78 сек.
ARIMA-LSTM	93.26% / 37.9%	9.82 сек.
ARIMA-GRU	93.01% / 37.9%	9.85 сек.

6. Заключение

Результаты тестирования на описанных наборах данных демонстрируют, что наш метод способствует не только повышению эффективности, но и уменьшению затрат времени исследователями. При все более широком использовании нейронных сетей настройка параметров является наиболее сложной и трудоемкой частью процесса. Поскольку наша модель в основном полагается на ARIMA, сокращение требований RNN позволяет экономить до 40-80% времени при решении той же проблеме. Это может помочь сэкономить время на настройку машинного обучения.

В целом, мы полагаем, что наша работа способствует формированию важного подхода в компьютерном прогнозировании временных рядов. Дальнейшая работа будет направлена на усиление модели и повышение ее производительности с экспериментальным тестированием на наборе данных GWAC, сокращению времени работы, особенно в процессе обучения нейронной сети.

Список литературы / References

[1]. V. Mayer-Schnberger, K. Cukier. Big data: A revolution that will transform how we live, work and think. Houghton Mifflin Harcourt, 2013, 256 p.

[2]. Sima Siami-Namini, Akbar Siami Namin, Forecasting Economics and Financial Time Series: ARIMA vs. LSTM, arXiv:1803.06386, 2018, 19 p.

[3]. G. Jenkins G.E.P. Box. Time series analysis, forecasting and control. Holden-Day, San Francisco, CA, 1970.

[4]. G.P. Zhang. Time series forecasting using a hybrid arima and neural network model. Neurocomputing, vol. 50, 2003, pp. 159-175.

[5] MIT-BIH Arrythmia Database: https://physionet.org/physiobank/database/mitdb/

Информация об авторах / Information about authors

Инь САН работает в Национальной лаборатории информатики и технологии, факультет компьютерных наук и технологий, Университет Цинхуа, Китай.

Ying SUN is working at Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, China.

Зиджун ЖАО работает в Национальной лаборатории информатики и технологии, факультет компьютерных наук и технологий, Университет Цинхуа, Китай.

Zijun ZHAO is working at Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, China.

Сяобин МА работает в Национальной лаборатории информатики и технологии, факультет компьютерных наук и технологий, Университет Цинхуа, Китай.

Xiaobin MA is working at Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, China.

Чжихуэй ДУ получил степень PhD в области компьютерных наук и технологий в Пекинском университете, КНР в 1998 г. В настоящее время он работает доцентом на факультете компьютерных наук и технологий университета Цинхуа, КНР. В число научных интересов входят параллельное программирование, высокопроизводительные, облачные, энергосберегающие вычисления и анализ больших данных.

Zhihui DU received the degree of PhD in Computer Science & Technology from Peking University, China in 1998. Currently he is the associate professor at the Department of Computer Science and Technology of Tsinghua University, China. His research interests include parallel computing, high performance/cloud/energy efficient computing, and Big Data analysis.