

Байесовский подход к обнаружению разладки

Татьяненко А.Д., СПбГУ, Санкт-Петербург alexdtat@gmail.com,
 Гориховский В.И., СПбГУ, Санкт-Петербург gorihovskyvycheslav@gmail.com,
 Кутуев В.А., Лаборатория технологий программирования инфраструктурных
 решений СПбГУ, Санкт-Петербург v.kutuev@spbu.ru

Аннотация

Рассматривается байесовский подход к обнаружению разладки. Описываются его преимущества для онлайн-задачи, позволяющие использовать алгоритм итеративно и локализовывать разладку по распределению длины пробега (количества шагов с последней разладки). Рассматриваются гауссовская и экспоненциальная предсказательные модели. Преимущество этих моделей на основе сопряжённых распределений — наличие явных формул для получения апостериорных гиперпараметров, что оказывается вычислительно простым решением. Предлагается эвристическая предсказательная модель, выбирающая на основе обучающей подвыборки более подходящую функцию правдоподобия из двух рассматриваемых. Описывается исследование зависимости этой эвристики от размера порога детектора и размера обучающей подвыборки, проведённое на различных распределениях. Основное внимание уделяется качеству локализации и задержкам.

Введение

Задача обнаружения разладки

Разладка — один из видов аномалий во временных рядах. Существуют различные определения, но в контексте данной работы под ней подразумевается точка смены одного распределения данных на другое (наблюдения на сегментах между ближайшими разладками независимы и одинаково распределены). Задача обнаружения разладки [3]:

$$\begin{aligned} &\{y_i\}, i = 1, \dots, n \\ &H_0: y_i \sim F_0, i = 1, \dots, n \\ &H_a: \exists \tau: 1 \leq \tau < n, y_i \sim \begin{cases} F_1, & i > \tau, \\ F_0 & \text{иначе.} \end{cases} \end{aligned}$$

Тогда τ — точка разладки. На практике также часто полезно не только обнаруживать разладку, но и определять момент времени, в который она произошла (при этом возможно некоторое отклонение от её настоящего местоположения), что далее будет называться *локализацией*. Эта задача возникает во многих контекстах, например, при анализе зашумлённых бенчмарков с целью определения момента падения производительности [4].

В данной работе рассматривается случай с размером данных $n = 500$ и разладкой (в случае её наличия) $\tau = 250$. Зависимость результатов работы алгоритма и подобранных значений параметров от размера данных и положения разладки здесь не рассматривается, являясь темой для отдельного исследования.

В связи с предоставлением небольшого набора алгоритмов в существующих Python-инструментах для обнаружения разладки и с отсутствием инструментов для сравнения соответствующих алгоритмов ведётся разработка PySATL-CPD. Рассматриваемый в данной статье алгоритм также входит в него.

Байесовский подход

В байесовском подходе к обнаружению разладки количество шагов, прошедших с последней разладки, рассматривается как случайная величина r_t (она называется длиной пробега) [2]. Её распределение можно представить как $P(r_t | \mathbf{x}_{1:t}) = P(r_t, \mathbf{x}_{1:t}) / P(\mathbf{x}_{1:t})$, где $\mathbf{x}_{1:t}$ — вектор наблюдений за время $1 : t$. Тогда, с учётом независимости данных, совместную вероятность для указанного выше распределения можно разложить следующим образом:

$$\begin{aligned} P(r_t, \mathbf{x}_{1:t}) &= \sum_{r_{t-1}} P(r_t, r_{t-1}, \mathbf{x}_{1:t}) = \\ &= \sum_{r_{t-1}} P(r_t, x_t | r_{t-1}, \mathbf{x}_{1:t-1}) P(r_{t-1}, \mathbf{x}_{1:t-1}) = \\ &= \sum_{r_{t-1}} P(r_t | r_{t-1}) P(x_t | r_{t-1}, \mathbf{x}_t^{(r)}) P(r_{t-1}, \mathbf{x}_{1:t-1}). \end{aligned} \quad (1)$$

В формуле 1 $\mathbf{x}_t^{(r)}$ означает, соответственно, наблюдения, относящиеся лишь к данным после последней зафиксированной разладки. Первый множитель (1) — это предсказательная вероятность, вычисляемая с помощью предсказательной модели. Более подробно они рассматриваются далее. Вторым множителем (1) определяется функция выживаемости. Она представляет априорное распределение на появление разладки, не зависящее непосредственно от наблюдений:

$$H(\tau) = \frac{P_{\text{gap}}(g = \tau)}{\sum_{t=\tau}^{\infty} P_{\text{gap}}(g = t)}, \quad (2)$$

где $P_{\text{gap}}(g)$ — дискретное априорное распределение на промежутке между разладками.

В зависимости от длины пробега указанная вероятность вычисляется следующим образом:

$$P(r_t | r_{t-1}) = \begin{cases} H(r_{t-1} + 1) & \text{если } r_t = 0 \\ 1 - H(r_{t-1} + 1) & \text{если } r_t = r_{t-1} + 1 \\ 0 & \text{иначе} \end{cases}$$

Как частный случай, если $P_{\text{gap}}(g)$ — геометрическое, то функция выживаемости (2) — константа: $H(\tau) = 1/\lambda$.

Последний множитель (1) оказывается рекуррентным, начиная с единственной возможной длины пробега, имеющей вероятность 1.0.

Адаптация алгоритма к онлайн-задаче

Онлайн-задача обнаружения разладки требует от алгоритма возможности обрабатывать данные по мере их поступления на протяжении долгого времени и по мере работы принимать решения об обнаружении разладки (алгоритм имеет доступ лишь к данным до текущего момента, а не ко всем сразу). Описанный выше подход хорош как математический метод, но требует адаптации для практического применения в случае решения онлайн-задачи.

Во-первых, требуется добавление явного критерия наличия разладки, поскольку в исходной работе [2] описано лишь получение распределения длины пробега. Для этого мной были выделены детектор, определяющий наличие разладки, и локализатор, определяющий её местоположение. В качестве детектора рассматривается порог вероятности максимальной длины пробега, поскольку именно она указывает на последнюю найденную разладку. В качестве локализатора предлагается выбор наиболее вероятной длины пробега из остальных.

Во-вторых, требуется возможность автоматически подбирать параметры предсказательной модели. Для этого в алгоритм был добавлен этап обучения предсказательной модели, на котором выделяется следующая за последней разладкой (или находящаяся в начале данных) подвыборка для последующей оценки на ней априорных гиперпараметров. Обучение после каждой разладки позволяет получать для каждого сегмента наблюдений более точную предсказательную модель.

В-третьих, требуется возможность работать достаточно долго. Исходный алгоритм имеет линейный рост требующихся на каждом шаге памяти и времени. Однако в случае принятия решения о наличии разладки можно удалять

все данные, относящиеся к сегменту ранее неё, поскольку распределения до и после разладки отличаются. В случае отсутствия разладки ситуация оказывается сложнее, поскольку приходится моделировать все возможные длины пробега, растущие линейно. Тем не менее, проанализировав характерную задержку при обнаружении разладки, можно предложить эвристику, основанную на параллельном запуске (с задержкой) и подготовке вспомогательного алгоритма, на который происходит переключение с основного по прошествии некоторого времени. Это будет приводить к потере информации, но позволит снизить сложность алгоритма до линейной вместо квадратичной на больших данных.

Предсказательные модели

Предсказательные модели позволяют оценивать правдоподобие принадлежности данных к распределению с оценёнными параметрами. В общем случае это требует интегрирования и решения оптимизационных задач, причём это оказывается необходимым для каждой длины пробега, поэтому количество таких операций на каждом шаге растёт линейно со временем. Общий вид предсказательной вероятности таков: $P(x_t | r_{t-1}, \mathbf{x}_t^{(r)})$, где $\mathbf{x}_t^{(r)}$ — вектор наблюдений для текущего пробега.

Хорошим частным случаем предсказательных моделей оказываются функции правдоподобия с сопряжёнными априорным/апостериорным распределениями. Они позволяют из обучающей подвыборки (обозначим её как \mathbf{x}^{learn}) оценивать априорные значения гиперпараметров, в явном виде вычислять их апостериорные значения из априорных с учётом полученного наблюдения (обозначим его как x^{obs}) и вычислять предсказательную вероятность с учётом прошлых наблюдений. Однако такие предсказательные модели покрывают достаточно небольшой класс распределений и потому требуют дополнительной оценки качества работы на данных, для которых не дают теоретических гарантий. Разные обозначения наблюдений из обучающей подвыборки и наблюдений, используемых при моделировании распределения длины пробега, вводятся для удобного разграничения разных этапов работы алгоритма. С формальной точки зрения они могут быть заменены просто наблюдениями с соответствующими временными индексами. Далее будут рассматриваться именно модели со свойством сопряжённости или эвристики на их основе.

Гауссовская предсказательная модель

При реализации базовой версии алгоритма (технически предназначенной для работы с данными фиксированного размера) была добавлена гауссовская предсказательная модель. С использованием сопряжённого нормального обратного-гамма распределения она позволяет оценивать и среднее, и дисперсию [5].

Экспоненциальная предсказательная модель

Гауссовская предсказательная модель показала себя заметно хуже на данных с использованием экспоненциального распределения или распределения Вейбулла. Хотя распределение Вейбулла и является обобщением экспоненциального распределения, для него не существует функции правдоподобия со свойством сопряжённости [6], поэтому было решено добавить экспоненциальную предсказательную модель. С использованием сопряжённого гамма-распределения она позволяет оценивать параметр экспоненциального распределения.

Эвристическая адаптивная предсказательная модель

Несмотря на то, что рассмотренные выше модели способны сами определять априорные значения параметров, на пользователе всё ещё остаётся ответственность за выбор самой модели. Кроме того, если рассматриваемые данные состоят из сегментов с распределениями из разных семейств, зафиксированная модель будет ограничивать алгоритм в гибкости. Поэтому было решено добавить эвристическую адаптивную модель, выбирающую на этапе обучения одну из функций правдоподобия (гауссовскую или экспоненциальную).

В качестве наивной эвристики было выбрано сравнение вероятностей получить обучающую подвыборку с учётом оценённых априорных гиперпараметров. Поскольку измерения предполагаются независимыми, при оценке вероятности для каждого измерения перемножаются. Стоит отметить, что корректный выбор предсказательной модели (например, с использованием критериев согласия) — отдельная большая задача, выходящая за пределы данной работы, но потенциально перспективная.

Численное исследование

Подготовка к исследованию

Для исследования алгоритма были сгенерированы синтетические данные с зафиксированными параметрами из следующих семейств: нормальное (до разладки $E[Y] = 0.0$, $\sigma = 1.0$, после разладки $E[Y] = 10.0$, $\sigma = 5.0$), равномерное ($E[Y] = 2.5$, $\sigma = 0.87$), бета ($E[Y] = 0.5$, $\sigma = 0.15$), экспоненциальное ($E[Y] = 0.2$, $\sigma = 0.2$) и Вейбулла ($E[Y] = 5.0$, $\sigma = 5.0$). Размер одного массива данных составлял 500, разладка (в случае её наличия) находилась в точке 250. Допустимое отклонение от размеченной (настоящей) разладки было выбрано $k = 25$ (параметр точности обнаружения разладки). Перебирались все возможные упорядоченные пары, а также случаи без разладки. Для каждой из конфигураций было сгенерировано по 1000 массивов данных.

На обоих этапах использовались постоянная функция выживаемости, пороговый детектор и локализатор, выбирающий наиболее вероятную (за исключением наибольшей, поскольку она соответствует уже найденной ранее разладке) длину пробега. Для функции выживаемости частота рассчитывалась так, чтобы априорная вероятность появления хотя бы одной разладки на 500 наблюдений составляла 0.5.

Гауссовская предсказательная модель

Было проведено численное исследование гауссовской предсказательной модели [1]. Сначала были определены значения порогов, соответствующие уровням значимости 0.1, 0.05, 0.01 и 0.005 на нормально распределённых данных (в Таблице 1 обозначено как α_N). На данном этапе исследования вычислялись распределения длины пробега на каждом шаге алгоритма, после чего проверялось первое пересечение порога. Это вело к тому, что ложно положительные (FP) срабатывания алгоритма исключали истинно положительные (TP). Далее были вычислены вероятности того, что при первом пересечении порога удавалось локализовать разладку в пределах допустимого отклонения от настоящей. В Таблице 1 эти вероятности обозначаются как $P_{Y_0-Y_1}$, где Y_0 — распределение до разладки, Y_1 — после.

В более чем 97% случаев на всём отрезке длины 500 где-либо была обнаружена разладка. При этом гауссовская предсказательная модель в ходе локализации разладки показала себя хуже при работе с экспоненциальным распределением, с распределением Вейбулла и с бета-распределением по сравнению с остальными случаями (Таблица 1). В случае экспоненциального распределения и распределения Вейбулла это может быть связано с их асиммет-

$\alpha_{\mathcal{N}}$	0.1	0.05	0.01	0.005
$P \backslash \text{Порог}$	0.27	0.18	0.07	0.04
$P_{\mathcal{N}-\mathcal{B}}$	0.794	0.824	0.864	0.878
$P_{\mathcal{N}-Exp}$	0.779	0.805	0.84	0.863
$P_{\mathcal{N}-U}$	0.87	0.893	0.917	0.929
$P_{\mathcal{N}-W}$	0.783	0.811	0.853	0.863
$P_{\mathcal{N}-\mathcal{N}}$	0.968	0.975	0.981	0.981

Таблица 1: Локализация с гауссовской предсказательной моделью

ричностью и существенно отличающимся носителем, из-за чего с точки зрения гауссовского правдоподобия многие наблюдения оцениваются как выбросы.

Эвристическая адаптивная предсказательная модель

При исследовании эвристической адаптивной предсказательной модели уже анализировалось поведение онлайн-алгоритма, а не базового, поэтому вместо оценки положения первой разладки оценивалось положение всех и попадание хотя бы одной из них в искомую окрестность. Были выбраны для оценки вычисленные ранее значения порогов, а также установлены размеры обучающих подвыборок 50, 20, 10 и 5 для предсказательных моделей.

Рассматривались данные такого же размера и с таким же положением разладки. Были выбраны следующие распределения с параметрами до разладки и после (если она была) соответственно:

- нормальное распределение \mathcal{N} с параметрами $\mu_0 = 1.0$, $\sigma_0 = 1.0$ и $\mu_1 = 10.0$; $\sigma_1 = 5.0$;
- равномерное распределение U с параметрами $a_0 = 0.0$, $b_0 = 1.0$ и $a_1 = 1.0$, $b_1 = 4.0$;
- бета-распределение \mathcal{B} с параметрами $\alpha_0 = 0.5$, $\beta_0 = 0.5$ и $\alpha_1 = 5.0$, $\beta_1 = 5.0$;
- экспоненциальное распределение Exp с параметрами $\lambda_0 = 1.0$ и $\lambda_1 = 5.0$;
- распределение Вейбулла W с параметрами $k_0 = 1.0$, $\lambda_0 = 0.5$ и $k_1 = 1.0$, $\lambda_1 = 5.0$.

Вероятность FP результата на данных без разладки оказалась не более 0.1 (случай нормального распределения и максимального порога, тогда как остальные вероятности FP оказались не более 0.06). В результате обнаружилось лишь 2 случая, где вероятность обнаружить разладку в пределах отклонения от настоящей (TP) была менее 0.79: $B - B$ и $W - U$. В обоих из них каждый раз на этапе обучения выбиралась экспоненциальная модель, тогда как для бета-распределения и для равномерного распределения лучше отработывала гауссовская модель. Предположительно, это связано с тем, что экспоненциальная модель выдавала большее правдоподобие для наблюдений из носителя $[0; 1]$, но хуже улавливала изменения плотности внутри него. Для решения этой проблемы, возможно, стоит рассмотреть ренормировку на этапе оценки предсказательной модели на обучающей подвыборке.

Также было замечено появление *дублирующих* ложно положительных (FP) разладок: на данных без разладок, как было уже упомянуто ранее, оказывалось сравнительно мало FP срабатываний, однако в случае наличия разладок количество FP срабатываний росло. При детальном рассмотрении оказалось, что это связано с некорректным выбором предсказательной модели после первой разладки, когда в обучающую подвыборку попадал сегмент данных до настоящей разладки, что мешало корректному выбору функции правдоподобия. Для устранения этого недостатка возможно использование эвристики, проводящих дополнительную обработку обучающей подвыборки (например, опуская первые m наблюдений).

Анализ задержек локализации показал, что, за вычетом оговоренных ранее плохих случаев, максимальная средняя из них составила 117, а максимальная медиана составила 110. Такие задержки, в совокупности с указанной выше вероятностью локализовать разладку с искомой точностью, указывают на возможность применения эвристики с параллельным вспомогательным алгоритмом, если выбирать общее рабочее время около 500 и время подготовки около 250.

Заключение

В работе была представлена адаптация байесовского алгоритма под решение онлайн-задачи обнаружения разладки¹. Далее были рассмотрены различные варианты предсказательных моделей, основанных на функциях правдоподобия со свойством сопряжённости: гауссовская, экспоненциальная и эвристическая адаптивная. Численное исследование на синтетических данных для гауссовской предсказательной модели показало работоспособность ал-

¹Исходный код (nick: alexdtat): <https://github.com/PySATL/pysatl-cpd> — 05.06.2025

горитма в случае использования гауссовской предсказательной модели. Эвристическая адаптивная модель частично позволила расширить область применения алгоритма, однако также появилось 2 случая, в которых она показывала себя плохо. Это связано с недостатками наивной эвристики, для борьбы с которыми в будущем можно рассматривать более продвинутые подходы к корректному выбору предсказательной модели. Также был приведён анализ вызванных эвристической моделью FP срабатываний алгоритма на данных с разладкой и предложен вариант решения проблемы попадания в обучающую подвыборку сегмента данных до настоящей разладки. Анализ задержек локализации указывает на возможность использования эвристики поддерживающего алгоритма для устранения роста затрат памяти и времени на каждом новом шаге алгоритма.

Список литературы

- [1] Татьяненко А. Адаптация и исследование байесовского подхода к обнаружению разладки для PySATL-CPD-Module. 2024. URL: [https://se.math.spbu.ru/thesis/texts/Tat'janenko_Aleksej_Dmitrievich_Spring_practice_3rd_year_2024_text.pdf](https://se.math.spbu.ru/thesis/texts/Tat%27janenko_Aleksej_Dmitrievich_Spring_practice_3rd_year_2024_text.pdf) — дата обращения 05.06.2025.
- [2] Adams R., MacKay D. Bayesian Online Changepoint Detection // Arxiv preprint. 2007. arXiv:0710.3742.
- [3] Aminikhanghahi S., Cook D. J. A survey of methods for time series change point detection // Knowl. Inf. Syst. 2017. T. 51. C. 339 –367.
- [4] Fleming M., Kolaczowski P., Kumar I. et al. Hunter: Using Change Point Detection to Hunt for Performance Regressions // Proc. of the 2023 ACM/SPEC International Conference on Performance Engineering (ICPE '23). 2023. C. 199 –206.
- [5] Murphy K. Conjugate Bayesian analysis of the Gaussian distribution. 2007.
- [6] Soland R. Bayesian Analysis of the Weibull Process With Unknown Scale and Shape Parameters // IEEE Trans. Reliab. 1969. T. R –18, № 4. C. 181 –184.