

АЛГОРИТМИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ЗНАЧИМЫХ ПРИЗНАКОВ ВСПЫШКИ КАК ИСТОЧНИКА СОЛНЕЧНЫХ ПРОТОНОВ

Ожередов В.А., Струминский А.Б.

Институт космических исследований РАН, Москва

ALGORITHMIC REVEALING OF THE SIGNIFICANT FEATURES OF SOLAR PROTON FLARE

Ozheredov V.A., Struminsky A.B.

Space Research Institute of RAS, Moscow

<https://doi.org/10.31725/0552-5829-2023-253-256>

Previously [1], it has been proposed that certain scripts of the solar flares dynamics show Solar Proton (SP) acceleration. We will refer to these flares as SP-sources. The proposed scripts were related to the dynamics of the flare temperature, its radio emission and the kinetic energy of the accompanying Coronal Mass Ejection (CME). In this paper, we suggest a machine learning method to identify a set of flare features that maximizes the Bayesian probability that a flare is an SP-source. In the general case, this problem resembles the well-known “Knapsack Problem”, the solution of which is impossible in a time polynomially depending on the size of input data (NP-hardness). However, to find the approximate solution of this task we developed so-called “search for parallel intersections”. This approach cut off inappropriate branches of the brute-force search strategy that allows us to obtain a criterion that is satisfied by 80% of SP-sources and only 1.1% of non-SP flares.

Следуя идеологии статьи [2], будем искать критерии SP-источника в виде объединения пересечений пороговых условий:

$$D = \bigcup_{k \in K} d^k = \bigcup_{k \in K} \bigcap_{\alpha=1}^m \left(\xi_{\alpha} \begin{matrix} \leq \\ \geq \end{matrix} y_{\alpha}^k \right) \quad (1)$$

Индекс k здесь нумерует элементарное решающее правило, срабатывающее при одновременном выполнении пороговых условий $\xi_{\alpha} \leq y_{\alpha}^k$ или $\xi_{\alpha} \geq y_{\alpha}^k$, где ξ_{α} – некоторый α -тый характерный признак вспышки, а y_{α}^k – его пороговое значение, поиск которого как раз и является задачей данной работы (понятно, что в общем случае порог зависит от номера рассматриваемого признака α и от индекса элементарного решающего правила k).

Немного необходимой для дальнейшего изложения терминологии. Радио-всплеском на заданной частоте мы будем называть событие, зафиксированное по данным RSTN (Radio Solar Telescope Network), радио-свипом мы называем радио-всплески II-го и IV-го типа, как это отмечено в файлах YYYYMMDDEvents.txt [3]. Каждому событию на временной оси соответствует начальная t_{begin} и конечная точка t_{end} . Радио-всплески и радио-сви́пы мы будем называть радио-событиями. Активной фазой вспышки мы называем максимальный по длине односвязный интервал, на котором температура превышает 12 МК. Будем говорить, что радио-событие на

интервале $[t_{begin}; t_{end}]$ относится к заданной вспышке, если этот интервал имеет непустое пересечение с ее активной фазой, см. рис. 1.

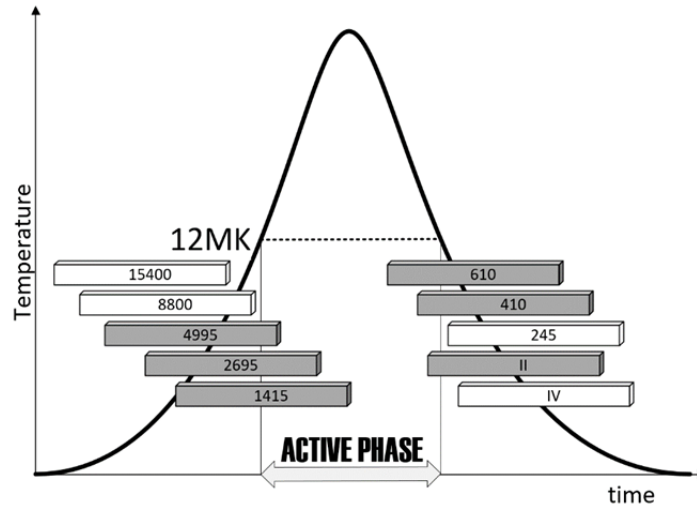


Рис. 1. Пересечение интервалов существования радио-событий с активной фазой вспышки. Признанные относящимися к вспышке радио-события помечены серым цветом.

Руководствуясь подходами статьи [1], в качестве характерных признаков вспышки для машинного обучения мы выбрали:

ξ_1 = число радио-всплесков на частотах 1415 МГц–15400 МГц, условно «красные» частоты;

ξ_2 = число радио-всплесков на частотах 245 МГц–610 МГц, условно «серые» частоты;

ξ_3 = наличие радиоизлучения II-IV-го типа безотносительно его типа.

Разумеется, перечисленные здесь радио-события должны относиться к заданной вспышке. Таким образом, у нас есть трехмерное пространство признаков, причем первый признак может принимать шесть значений, от 0 (нет соответствующих частот) до 5 (все частоты присутствуют), второй признак – четыре значения, от 0 до 3, и третий признак – два значения, поскольку это битовая величина. Таким образом, множество альтернатив

$$\Omega = \left\{ \omega^i \equiv \bigcap_{\alpha=1}^m (\xi_{\alpha} = x_{\alpha}^i) \right\} \quad (2)$$

содержит $2^{6+4+2} = 281\,474\,976\,710\,656$, то есть примерно триста миллиардов элементов, которые исчерпывают все мыслимые решающие правила.

Пусть C обозначает событие (в вероятностном смысле) «**заданная вспышка является SP-источником**», а D – некоторое решающее правило, т.е. событие «**характерные признаки вспышки удовлетворяют объединению пересечений неравенств (1)**». Понятно, что наилучшее решающее правило одновременно максимизирует как условную вероятность появления SP-источника при наличии признаков, заданных правилом D , так и ве-

роятность появления этих признаков при условии, что вспышка является SP-источником. Первое условие минимизирует false positive rate, т.е. вероятность принять обычную вспышку за SP-источник, а второе – false negative rate, или вероятность не распознать истинный SP-источник. По формуле Байеса

$$\begin{aligned} P(C|D) &= \frac{P(C \cap D)}{P(D)} = \frac{P(D \cap C)}{P(D \cap C) + P(D \cap \bar{C})} = \\ &= \frac{1}{1 + \frac{P(D \cap \bar{C})}{P(D \cap C)}} = \frac{1}{1 + \frac{P(\bar{C})}{P(C)} \frac{P(D|\bar{C})}{P(D|C)}} = \frac{1}{1 + \gamma \frac{P(D|\bar{C})}{P(D|C)}} \end{aligned} \quad (3)$$

(здесь γ это не волнующий нас множитель, равный соотношению частот появления SP-источников и обычных вспышек, а горизонтальная черта над событием означает его *не*-появление). Поскольку, как видно из формулы (3), искомая вероятность монотонно растет с уменьшением обратного отношения правдоподобия $L^{-1}(D)$, будем иметь следующую оптимизационную задачу:

$$\begin{cases} L^{-1}(D) = \frac{P(D|\bar{C})}{P(D|C)} \rightarrow \min \\ P(D|C) \rightarrow \max \end{cases} \quad (4)$$

Проблема таким образом сводится к перебору всевозможных комбинаций альтернатив из Ω , см. формулу (2), что примерно соответствует известной задаче «knapsack problem» [4]: отправляясь в поход, необходимо набрать себе в рюкзак еды из холодильника, содержащего примерно триста миллиардов наименований продуктов (напомним, что примерно столько элементов содержит в себе множество 2^Q), так, чтобы поход был минимально тяжелым и максимально сытным. Доказано, что эта задача является NP-трудной, т.е. алгоритмически неразрешимой за полиномиально зависящее от входных данных время – отсюда NP = Non-deterministic Polynomial.

Мы показали, что существует упорядочение альтернатив из множества (2)

$$\omega^{(1)}, \omega^{(2)}, \dots, \omega^{(6 \cdot 4 \cdot 2)} \quad (5)$$

такое, что обратное правдоподобие от куска «вариационного» ряда (5)

$$f(n) \equiv L^{-1} \left(\bigcup_{i=1}^n \omega^{(i)} \right)$$

– монотонная, но при этом медленно растущая функция. Это означает возможность поиска пороговых решающих правил d^k , см. формулу (1), при помощи итеративного алгоритма, который мы назвали методом параллельных пересечений. Заметим, что пороговых решающих правил достаточно немного: если переменная ξ_α принимает всего m значений, то поро-

говых правил только лишь с ее участием насчитывается $2m - I$, откуда следует, что общее число участвующих в задаче пороговых решающих правил равно $(2 \cdot 6 - I) (2 \cdot 4 - I) (2 \cdot 2 - I) = 23I$. Отсюда вытекает возможность итеративного объединения пороговых решающих правил, где на каждой итерации I отбираются такие $d^k(I)$, объединение которых полностью покрывает кусок вариационного ряда

$$\bigcup_{i=1}^I \omega^{(i)}$$

При этом false positive rate будет расти, а false negative rate – уменьшаться, и алгоритм следует остановить, когда false positive rate превысит заранее определенную величину, например, 1%.

Мы получили список SP-источников, сопоставив вспышки из таблицы [5] со списком ~180 000 электромагнитных событий из базы данных [3]. Применяв описанный выше алгоритм, мы получили решающее правило:

если в активной фазе имеется не менее 2 красных частот, не менее 3 серых частот **и** есть излучение II-IV типа,

или не менее 4 красных частот, не менее 1 серой частоты **и** есть излучение II-IV типа,

или не менее 5 красных частот **и** не менее 2 серых частот,

то вспышка скорее всего сгенерирует протонное событие.

Из 55-ти принявших участие в эксперименте протонных событий правильно классифицировано 80%, а из ~27 000 обыкновенных вспышек неправильно классифицированы 1.1%. Средняя энергия протонов, вычисленная по данным каталога [6], из правильно классифицированных SP-источников равна 441 МэВ, а неправильно классифицированных – 136 MeV. При этом ручной анализ неправильно классифицированных SP-источников показал, что добавление к пространству признаков еще одной переменной – скорости сопровождающего вспышку коронального выброса массы >900 км/с, позволяет свести false negative rate к нулю.

Литература

1. Струминский А.Б., Садовский А.М., Григорьева И.Ю. Критерии для предсказания протонных событий по солнечным наблюдениям в реальном времени // По материалам симпозиума Прогноз-2023, ИЗМИРАН, подана в редакцию G&A, 2023.
2. Swalwell B., S. Dalla, R.W. Walsh. Solar Energetic Particle Forecasting Algorithms and Associated False Alarms // Solar Physics, 2017
3. База данных по электромагнитным событиям:
https://cdaw.gsfc.nasa.gov/CME_list/NOAA/org_events_text/
4. Описание задачи о рюкзаке: https://en.wikipedia.org/wiki/Knapsack_problem
5. Список SP-источников по версии NASA: <https://umbra.nascom.nasa.gov/SEP/>
6. Ссылка на каталог SP-источников по версии НИИЯФ:
https://swx.sinp.msu.ru/apps/sep_events_cat/docs/SPE_24_Summary_List.pdf