

Построение распределения времени жизни солнечных пятен с помощью алгоритмов машинного обучения

Садртдинов Ильдус

irsadrtdinov@edu.hse.ru

Предисловие. На сегодняшний день проведено несколько исследований, посвященных изучению времени жизни солнечных пятен. В этой научно-исследовательской работе использован кардинально новый подход к этой задаче - площадной метод на основе метрического алгоритма k-ближайших соседей.

Ключевые слова: Солнечные пятна, солнечная активность, время жизни солнечных пятен, распределение вероятности, метод k-ближайших соседей, параметрическое моделирование

1 Введение

База данных предоставлена Гринвичской королевской обсерваторией (The Royal Greenwich Observatory (RGO)). В ней содержится информация о наблюдаемых солнечных пятнах в период с мая 1874 года по декабрь 1974, а именно: дата и время наблюдения, наблюдаемая и скорректированная (с учетом расположения на солнечной сфере) площадь солнечного пятна (измеряемая в миллионных долях площади солнечной полусферы), а так же долгота и широта пятна в Кэррингтоновских гелиографических координатах.

Основная проблема, связанная с построением распределения времени жизни солнечных пятен, состоит в том, что наблюдения Солнца ведутся с Земли, поэтому одновременно можно получить информацию только о той половине солнечной сферы, что обращена к Земле. Если пятно существует больше половины периода обращения Солнца, то оно уходит из поля зрения наблюдателя, и дальнейшее его развитие происходит на невидимой стороне. По этой причине в каталоге нет пятен, которые наблюдались больше 15-16 дней (примерно половина периода обращения Солнца около полюсов). Кроме того, в базе данных получается неестественно много пятен с временем жизни в 12-13 дней, что видно на рис. 1, поскольку история долгоживущих пятен «обрезается». В итоге, основной задачей научно-исследовательской работы является устранения «горба» на графике распределения времени жизни и построение правдоподобного распределения, в котором доля пятен, существующих 14 дней и больше, была бы отлична от нуля, и доля убывала бы с ростом времени жизни.

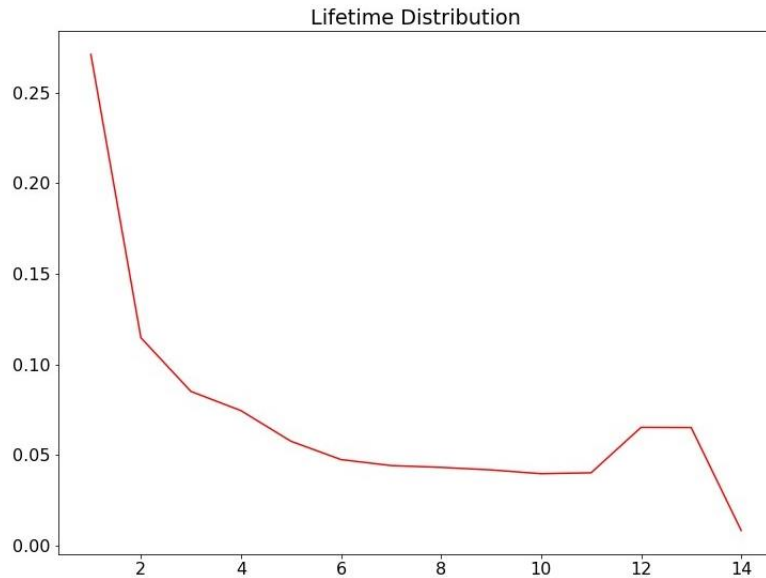


Рис. 1. Распределение времени жизни солнечных пятен на основе исходных данных. По горизонтальной оси отложены дни, по вертикальной оси – соответствующая им доля пятен.

2 Алгоритм предсказания площади

Площадь солнечного пятна – одна из основных его характеристик, которая меняется вместе с развитием пятна. Прежде всего был разработан алгоритм, предсказывающий площадь пятна на один день (следующий после имеющийся истории). Алгоритм основан на применении регрессионного метода к-ближайших соседей.

Введем следующие обозначения: под S^i будем понимать некоторое солнечное пятно, а под S_j^i – площадь пятна S^i в j -ый день его наблюдения. Для применения метода k -ближайших соседей необходимо определить метрику на множестве солнечных пятен. Рассмотрим евклидово пространство \mathbb{R}^c , где c – некоторый параметр. В качестве элементов этого пространства будем рассматривать c -мерные векторы $(S_i^1, \dots, S_{i+c-1}^1)$ для всех возможных пятен S и всех индексов i , для которых указанные площади известны при фиксированном S . В качестве метрики на \mathbb{R}^c возьмем стандартное евклидово расстояние, то есть для $S^1 = (S_i^1, \dots, S_{i+c-1}^1)$ и $S^2 = (S_j^2, \dots, S_{j+c-1}^2)$:

$$\rho(S^1, S^2) = \sqrt{\sum_{k=0}^{c-1} (S_{i+k}^1 - S_{j+k}^2)^2} \quad (1)$$

Таким образом, расстояние между двумя пятнами будет зависеть от дней, относительно которых оно считается. Будем считать расстояние между пятнами S^* , S относительно дней d^* , d как:

$$\rho(S^*, S, d^*, d) = \sqrt{\sum_{k=0}^{c-1} (S_{d^*-k}^* - S_{d-k})^2} \quad (2)$$

То есть так, чтобы значения площади S^* , S в дни d^* , d являлись последней координатой соответствующего c -мерного вектора. На практике это означает, что эти дни будут последними днями истории для учета при подсчете расстояния.

Итак, вернемся к исходной задаче предсказания площади. Пусть для пятна S^* доступна история в N дней (в частности, известны значения площадей S_1^* , ..., S_N^* в каждый из этих N дней). Необходимо приписать ему значение площади в $N+1$ день. Для этого выберем k (k – параметр модели) ближайших соседей в смысле расстояния, заданного формулой (2). Для этого нужно положить $d^* = N$ (последний день истории пятна S^*). Пусть пятно S – кандидат в ближайшие соседи – имеет историю в M дней. Тогда значение d для пятна S неизвестно. Поэтому чтобы вычислить оптимальное расстояние между S^* и S необходимо минимизировать расстояние ρ из (2) по d :

$$\rho_{opt}(S^*, S, d^*) = \min_{d \in [c, M]} \rho(S^*, S, d^*, d) \quad (3)$$

Далее, относительно ρ_{opt} нужно найти k ближайших соседей. Пусть это будут пятна S^1, \dots, S^k с оптимальными днями d_1, \dots, d_k (в которых достигается минимум ρ) и расстояниями ρ_1, \dots, ρ_k . В итоге, значение площади пятну S^* приписывается следующим образом (среднее арифметическое площадей соседей):

$$S_{N+1}^* = \frac{1}{k} \sum_{i=1}^k S_{d_i+1}^i \quad (4)$$

Значения параметров были выбраны следующими:

$$c = 3, k = 3 \quad (5)$$

Кроме того, площадь можно приписывать как взвешенное среднее площадей соседей (этот способ показал себя как более эффективный), а именно:

$$S_{N+1}^* = \sum_{i=1}^k w_i S_{d_{i+1}}^i \quad (6)$$

При этом, веса (вклады площадей соседей) вводились обратно пропорционально расстояниям до соответствующих пятен, то есть:

$$\begin{cases} \frac{w_i}{w_j} = \frac{\rho_j}{\rho_i}, 1 \leq i < j \leq k \\ \sum_{i=1}^k w_i = 1 \end{cases} \Leftrightarrow w_i = \frac{1}{\rho_i \sum_{j=1}^k \frac{1}{\rho_j}} \quad (7)$$

Далее проводилось тестирование алгоритма. Относительная ошибка вычислялась по формуле:

$$\varepsilon = \frac{S_{pr} - S}{S} \quad (8)$$

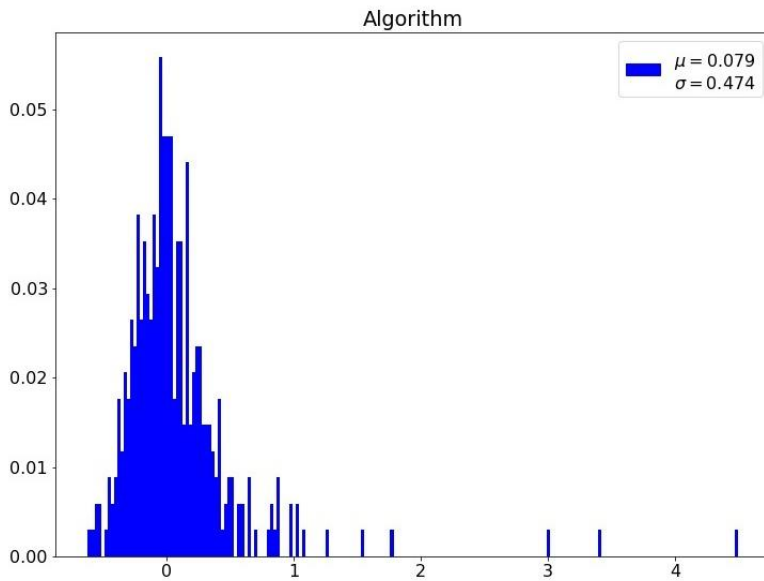


Рис. 2. Распределение ошибок алгоритма (среднего арифметического). По горизонтальной оси отложена ошибка, по вертикальной – соответствующая ей доля. На графике μ – среднее (арифметическое) ошибки, σ – стандартное отклонение.

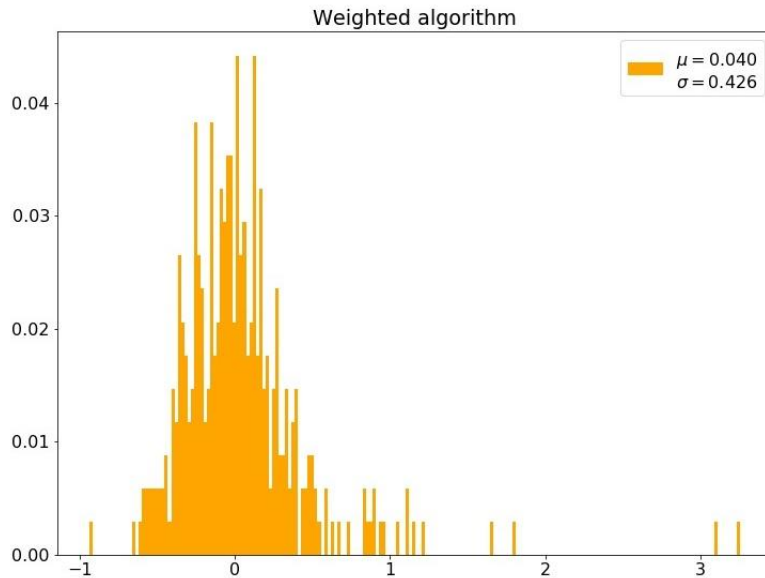


Рис. 3. Распределение ошибок взвешенного алгоритма.

На рис. 2 – 3 изображены распределения ошибок, которые выдают алгоритмы на одной и той же выборке из 1000 пятен. Однако, на этих графиках не учтены ошибки, получающиеся на пятнах с площадью, меньшей $S_{min} = 120$, поскольку для пятен с относительно небольшой площадью свойственные резкие («скачко-вые») уменьшения площади, на которые метрический алгоритм не реагирует, что задает границы применимости алгоритма. Это свойство будет в дальнейшем использовано для моделирования площади пятна на дистанции.

3 Симуляция жизни солнечных пятен

Имея в распоряжении алгоритм, предсказывающий площадь на следующий день, можно приписывать площадь на несколько дней вперед, запуская его последовательно. Тем не менее, необходимо отдельно симулировать скачковое изменение площади, которое игнорирует алгоритм. Для этого следующим образом была построена модель с параметрами $\alpha \in (0,1), p \in [0,1]$.

- С вероятностью p площадь алгоритма уменьшается в $1/\alpha$ раз, то есть:

$$S_{N+1}^* = \alpha \cdot S_N^* \quad (9)$$

- Иначе (с вероятностью $1 - p$) площадь приписывается согласно взвешенному алгоритму.
- Если площадь пятна становится меньше значения $S_{lt} = 20$, то пятно считается «умершим» и дальнейшее приписывание площади для него не ведется.

На основе поведения настоящих солнечных пятен были выбраны значения параметров:

$$\alpha = \frac{1}{3}, p = \frac{1}{7} \quad (10)$$

4 Распределение времени жизни

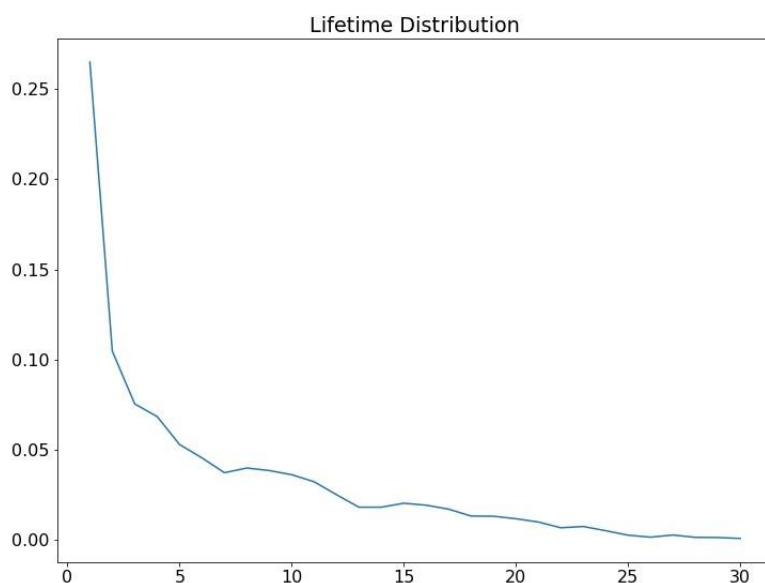


Рис. 4. Скорректированное алгоритмом распределение времени жизни солнечных пятен. По горизонтальной оси отложены дни, по вертикальной оси — соответствующая им доля пятен.

На основе модели было построено распределение времени жизни, изображенное на рис. 4. Алгоритм применялся для пятен, история которых обрывалась на правом краю Солнца, а потому их настоящее время жизни было неизвестно. В результате симуляции история солнечного пятна дополнялась, продлевая время жизни пятна. Для остальных пятен время жизни бралось из исходных данных. Применение алгоритма позволило «сгладить горб», который появлялся на графике распределения, построенном по исходным данным. В итоге, полученное распределение времени жизни солнечных пятен, достоверность которого опирается на точность алгоритма предсказания площадей и естественные процессы развития солнечных пятен.

СПИСОК ИСТОЧНИКОВ

1. Truls S. Ringnes. On the lifetime of sunspot groups / Institute of Theoretical Astrophysics, University of Oslo // *Astrophysica Norvegica* – 1964. – Vol. IX, No. 11. P.95-101.
2. Astropy Documentation – Astropy v3.1 [Электронный ресурс] / The Astropy Developers. URL: <http://docs.astropy.org/en/stable/index.html>, свободный. (дата обращения: 16.12.18).
3. Coordinate systems for solar image data [Электронный ресурс] / W. T. Thompson ESO 2005. URL: http://www.mssl.ucl.ac.uk/grid/iau/extra/local_copy/coordinates.pdf, свободный. (дата обращения 19.03.2018)
4. Matplotlib Documentation – Matplotlib version 3.0.2 [Электронный ресурс] / 2002 - 2012 John Hunter, Darren Dale, Eric Firing, Michael Droettboom and the Matplotlib development team; 2012 - 2018 The Matplotlib development team. URL: <https://matplotlib.org>, свободный. (дата обращения: 15.02.19).
5. NumPy [Электронный ресурс] / 2018 NumPy developers. URL: <http://www.numpy.org>, свободный. (дата обращения: 15.02.19).
6. Random – Generate pseudo-random numbers [Электронный ресурс] / 2001-2019, Python Software Foundation. URL: <https://docs.python.org/3/library/random.html>, свободный. (дата обращения: 19.03.2019)
7. Royal Observatory, Greenwich USAF/NOAA Sunspot Data [Электронный ресурс] / Dr. David Hathaway. URL: <https://solarscience.msfc.nasa.gov/greenwch.shtml>, свободный. (дата обращения: 16.12.18).
8. SunPy Documentation – SunPy v0.9.5 [Электронный ресурс] / The SunPy Community. URL: <http://docs.sunpy.org/en/stable/guide/index.html>, свободный. (дата обращения: 16.12.18).
9. The Lifetime of a Sunspot Group [Электронный ресурс] / Commonwealth of Australia 2019, Bureau of Meteorology. URL: <http://www.sws.bom.gov.au/Educational/2/2/2/>, свободный. (дата обращения: 20.03.2019).
10. The Sun's Vital Statistics [Электронный ресурс] / 2008 Stanford SOLAR Center. URL: <http://solar-center.stanford.edu/vitalstats.html>, свободный. (дата обращения: 19.03.2019).