

Моделирование стихийных бедствий

Коробкина Ю.Ш., ООО «Синкретис», директор по развитию,
г. Томск, Россия

SPIN-код: 4649-4746

Осипов А.В., к. физ.-мат. н., ООО «Синкретис»,
ст. математик-программист, г. Санкт-Петербург, Россия

SPIN-код: 6217-7670

Шендеров В.Ю., ООО «Синкретис», инженер данных,
г. Томск, Россия

SPIN-код: 3100-6718

Аннотация

В статье обсуждается ряд моделей стихийных бедствий, входящих в состав продукта Smart Geo Platform (SGP) компании ООО «Синкретис», а именно: модели детекции пожаров, прогнозирования вероятности пожаров, распространения лесных пожаров и распространения наводнений.

Ключевые слова: машинное обучение; моделирование; предотвращение; детекция; распространение; пожары; наводнения.

Введение и постановка задачи

В статье приводится описание нескольких моделей стихийных бедствий, входящих в продукт SmartGeoPlatform (SGP) компании ООО «Синкретис» (запись в реестре Российского ПО от 01.03.2023 № 16764). А именно: будут представлены задачи моделирования пожаров и наводнений.

Обсуждение моделей пожаров будет построено следующим образом. Сначала будет описано решение задачи детекции пожаров и гарей по спутниковым снимкам. Далее будет представлена модель прогнозирования пожаров (т.е. определения вероятности нахождения пожара в данной точке). И, наконец, будет показана модель распространения лесного пожара. Неформально задачу моделирования лесного пожара

можно описать следующим образом: обнаружить пожар на снимке и определить возможную область распространения за определенное время, т.е. ответить, например, на вопрос: может ли пожар за 6 часов добраться до поселений?

После представления модели по пожару будет описана модель наводнения. Эту задачу можно сформулировать так: есть снимок, на котором отмечена река, необходимо смоделировать разлив реки при наводнении. Наводнения бывают двух основных типов: сезонные и паводковые. По сезонным наводнениям, как правило, есть исторические данные, и для них можно применять подходы, основанные на машинном обучении, тогда как паводковые наводнения происходят внезапно, и по ним таких данных нет. Представленная в этой работе модель посвящена паводковым наводнениям.

Модель детекции пожаров

Для задачи детекции пожаров существует несколько общепринятых подходов. Большая часть из них опирается на индексы со спутников Sentinel или Landsat. Индексом называется некая алгебраическая комбинация каналов снимка дистанционного зондирования Земли (ДЗЗ). В статье [5] проводятся детальный анализ и сравнение качества работы разных индексов с наглядной демонстрацией, какие плюсы и минусы есть у каждого из них.

Для модели детекции пожаров были выбраны индексы NBR (Normalized Burned Ratio), dNBR (deltaNBR), NDSWIR, BAI (Burned Area Index), а также их комбинации и эвристически-определенные пороговые значения.

Выбранный подход определения выгоревших зон использует объединение трех масок:

1. Маска по пороговому значению для индекса NBRSWIR $(swir1 - swir2)/(swir1 + swir2)$.

2. Маска по пороговому значению для индекса NDVI $(nir + red)/(nir - red)$.

3. Маска по пороговому значению для индекса NBR $(swir2 - nnir)/(nnir + swir2)$.

Далее, объединение этих масок проходит через суммирующий фильтр размером 7×7 . Алгоритм работы напоминает таковой для

операции `max_pooling` (см. рис. 1), где вместо максимума считается сумма по всем пикселям окна. Далее, если значение пикселя-суммы превышает пороговое, то считаем, что в данном пикселе тоже произошел пожар.

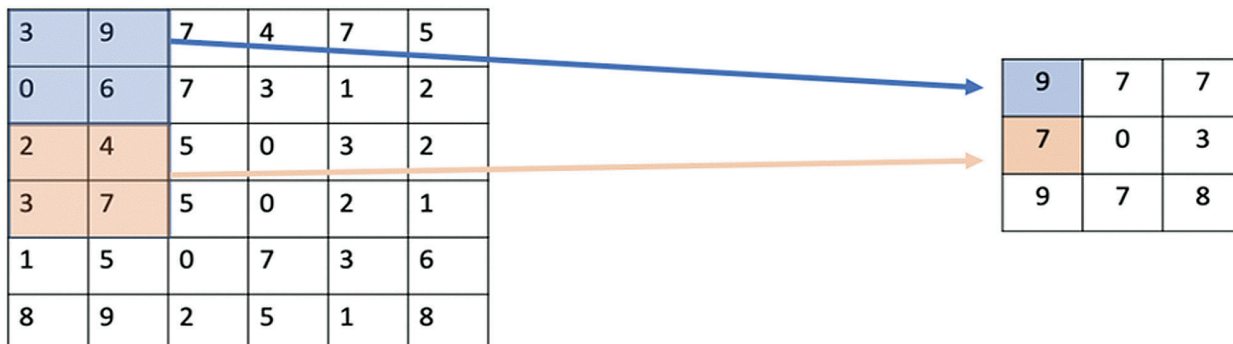


Рис. 1. Описание работы слоя `max_pooling`

Также модель умеет определять непосредственные места пожара (активные очаги), для чего используются канал `SWIR2` и схожая маска со скользящим окном.

На рис. 2 представлены результаты работы модели детекции пожаров.



Рис. 2. Результаты работы модели детекции пожаров

Как видно, модель детекции пожаров работает достаточно хорошо.

Модель предсказания пожаров

Модель разработана на основе соревнования `Predicting Wildfires with Weather Forecast Data` [8]. Тренировочный датасет также был взят с этого соревнования, потому что использует данные того же спутника,

что доступны на сервисе Open-Meteo [13]. Для решения этой задачи необходимо агрегировать почасовые данные о погоде в ежедневные с выделением среднего, минимального, максимального значений, а также среднеквадратичного отклонения.

Итоговый список параметров, используемых для обучения модели и предсказания пожаров (в скобках указаны способы агрегации для каждого из параметров): <Temperature>:[‘mean’, ‘min’, ‘max’, ‘var’]; <RelativeHumidity>:[‘mean’, ‘min’, ‘max’, ‘var’]; <WindSpeed>:[‘mean’, ‘min’, ‘max’, ‘var’]; <Precipitation>:[‘sum’]; <SoilWaterContent>:[‘mean’, ‘min’, ‘max’, ‘var’]; <SolarRadiation>:[‘sum’]

После того как параметры корректно агрегированы в дневные значения, необходимо рассчитать скользящие статистики для каждого из значений. Для нашей задачи выбраны скользящие окна шириной 7 и 3 дня. При этом для каждого из параметров рассчитывается среднее, а для среднего и суммы рассчитывается еще и среднеквадратичное отклонение. Таким образом, для каждого из параметров (из списка выше) генерируется дополнительно по 2–3 параметра. Итого, для обучения модель получает на вход 66 параметров.

В качестве классификатора модель использует SGDClassifier с logloss, что дает в результате модель логистической регрессии. Дополнительно устанавливается “L2” регуляризация, чтобы бороться с неравномерностью коэффициентов в уравнении логистической регрессии. Дополнительным преимуществом модели SGD является возможность «дообучения», когда модель может не просто учиться на фиксированном наборе данных, а как и нейронные сети дообучаться в процессе. Это может позволить модели совершенствоваться и лучше подстраиваться под данные конкретного региона со временем. Используемая модель схожа с моделью, описанной и протестированной в [12].

В итоговом обучающем датасете содержится 40 тысяч строк, отображающих погоду в разных регионах в течение последних нескольких лет. По записям о пожарах данные о погоде были размечены следующим образом. Если в какой-то день зафиксирован пожар, то этот день и два предшествующих дня помечаются как дни, в которых зафиксирован пожар. Итоговый баланс классов в выборке составил около 0.65. При обучении использовался подход бэггинга, когда отдельные батчи генерировались с равным балансом классов и отправлялись в обучение модели.

Результаты работы обученной модели на тестовых данных: Holdoutsetresults: acc: 0.73 roc_auc: 0.648 f1: 0.8 classbalance: 0.6.

На рис. 3 представлены параметры, оказавшие максимальный эффект на принятие решения моделью.

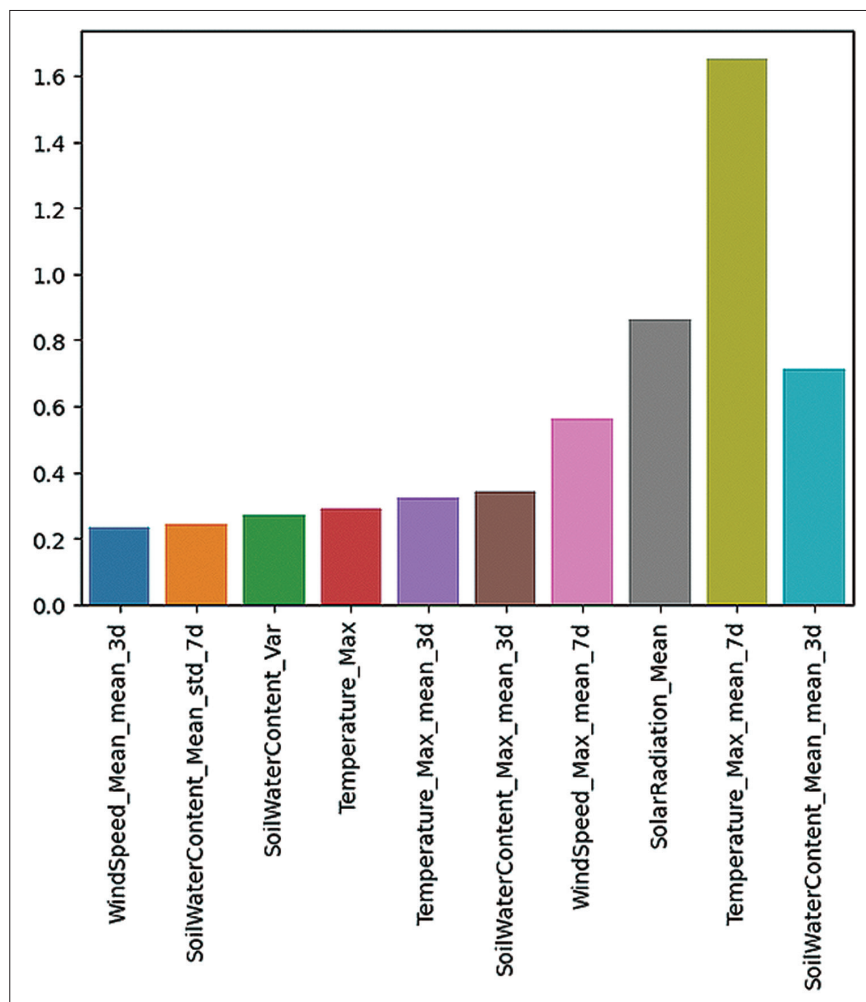


Рис. 3. Наиболее значимые признаки

Как видно из рисунка 3, параметры вполне соответствуют интуитивному пониманию условий, когда должен произойти пожар. А именно, максимально жаркие дни (за последнее время, 7 дней в данном случае) с наибольшим количеством солнечной радиации (безоблачная погода).

Модель распространения пожаров

Существует достаточно много публикаций на тему распространения лесных пожаров. Классической моделью является модель Rothermel [14]. Существует множество модификаций этой классической модели

[17, 16, 6]. Есть известные модели, не являющиеся вариациями модели Rothermel: модель Alexander&Cruz — для пожара лесной кроны [1] и модель CFFBR, имеющая другие параметры [7]. Кроме того, есть достаточно перспективный подход, основанный на применении клеточных автоматов [2].

На вход модель распространения принимает снимок с размеченным пожаром. Входные данные включают в себя:

- момент времени, на который осуществляется прогнозирование;
- скорость ветра (вектор);
- температуру;
- тип деревьев;
- базальную площадь деревьев;
- высоту деревьев;
- плотность деревьев;
- время, в течение которого возникает местное возгорание;
- 100-часовую загрузку;
- глубину топливного слоя;
- топографический наклон.

Эти данные могут меняться со временем и быть разными для разных точек снимка. Все эти неоднородности поддерживаются моделью.

На выходе модель выдает область максимального распространения вокруг каждого участка пожара для этого момента времени. Речь идет о максимальном распространении, так как важен именно пессимистичный прогноз.

Модель начинает работу с обнаружения очагов пожара. Задача нахождения очагов — это задача кластеризации. Существует множество алгоритмов кластеризации. Предлагается использовать алгоритм HDBSCAN [4] из библиотеки `dbscan (R)`. Авторы этого алгоритма провели множество экспериментов по сравнению его с другими доступными. Они заключили, что их алгоритм лучше в определенном смысле, вдобавок он сам умеет определять число кластеров по заданному минимальному размеру кластера.

Распространение лесных пожаров проводится в основном с помощью различных функций открытого пакета `firebehavioR (R)`. Функция `canFuel` из пакета `firebehavioR`, ориентированного на случай Канады, поддерживает только 4 типа леса (смешанный лес, желтая сосна, скрученная сосна, псевдотсуга Мензиса). Так как тип леса влияет на

распространение пожара, было решено обобщить функцию `canFuel` для работы с еще несколькими десятками типов леса, встречающихся в России. Эта функция для каждого типа леса возвращает высоту и плотность лесной кроны. Идея русификации состоит в том, чтобы разработать модели для встречающихся в России типов леса, исходя из открытых данных. На основе изучения различных открытых источников данных определялись средняя высота и диаметр кроны для каждого типа деревьев. На приведенном ниже рис. 4 показаны позиции различных российских деревьев относительно деревьев из `firebehaviorR`. Красным обозначены деревья из `firebehaviorR`. Разными цветами выделены разные кластеры деревьев, которые по-разному подвергаются интерполяции или экстраполяции.

Приведем здесь итоговый состав кластеров (слева направо на рис. 4).

Внешний кластер. Можжевельник, тис, рябина.

Кластер 1. Бук, липа, ольха, граб, клен, ива.

Кластер 2. Сосна, кедр, дуб, ясень, вяз, береза, осина.

Кластер 3. Лиственница, ель, пихта, тополь.

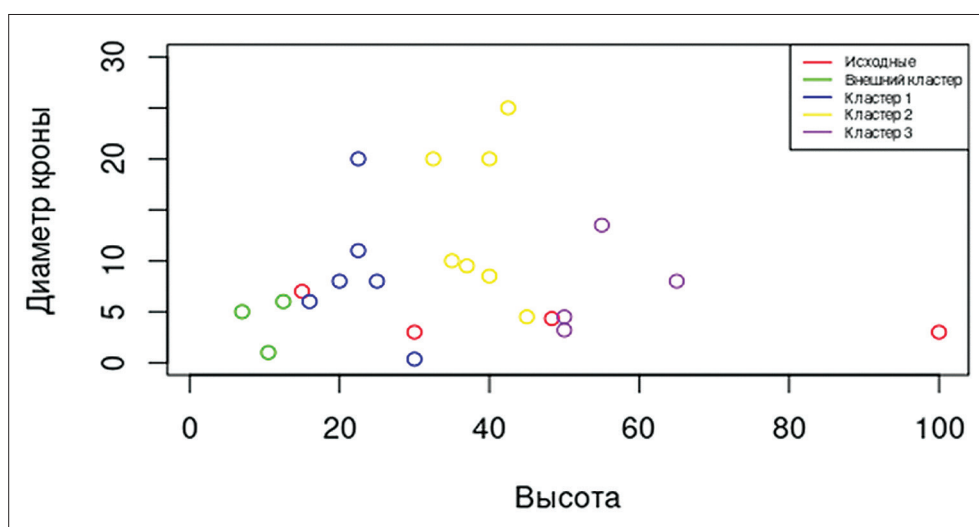


Рис. 4. Кластеризация типов деревьев

Лесные пожары бывают 3 типов:

- поверхностные (`surfacefires`), которые распространяются сравнительно медленно, огонь сравнительно невысок;
- лесной кроны (`crownfires`), они хоть и встречаются реже поверхностных, но распространяются гораздо быстрее;

- местные (spotfires), пожары, возникающие вдали от основного пожара, если этот пожар является пожаром кроны, причиной может быть, например, вылетевшая горящая шишка.

В качестве модели пожара кроны взята модель Alexander&Cruz [1], реализованная в качестве функции `sfis` из пакета `firebehavioR` (R). Авторы статьи сравнивали свою модель с другими моделями и заключили, что их модель выдает наиболее пессимистичные прогнозы. Они считают это несомненным преимуществом модели из-за специфики применения модели распространения лесных пожаров. Модель также поддерживает местные возгорания (например, шишки, которые вылетают и чуть-чуть увеличивают основной фронт пожара). Это эмпирическая модель, позволяющая оценить скорость распространения пожара и вероятность пожара лесной кроны по данным 6 параметров (количество сожженного топлива, скорость ветра, влажность, разрывы топливного слоя, плотность лесной кроны, время, через которое возникает местное возгорание).

В качестве модели поверхностных пожаров используется вариация модели `Rothermel`, реализованная в функции `rothermel` из пакета `firebehavioR`, (R). Речь идет об эмпирической модели [16]. Эта модель очень гибка, однако у нее 30 входных параметров, что, конечно, слишком много. Поэтому было решено провести анализ чувствительности модели `Rothermel`, выделив наиболее значимые параметры, а остальные заменив на разумные значения по умолчанию. Анализ чувствительности основывался на открытых данных: датасеты `fuel Models` и `fuel Moisture` из пакета `firebehavioR` (R). В ходе анализа чувствительности было выбрано 6 параметров, наиболее важных для модели (100-часовая загрузка, глубина топливного слоя, количество поглощенного топлива, наклон, скорость ветра, поправка на ветер). Таким образом, модели пожаров лесной кроны и поверхностных пожаров были приведены к единому знаменателю: у каждой из них стало по 6 параметров, правда эти параметры разные.

Вероятность пожара лесной кроны оценивается с помощью модели Alexander&Cruz, дальше берется взвешенная сумма результатов двух описанных выше моделей с весом, привязанным к вероятности. После этого получаем модельную скорость распространения пожара. Пользователю предоставляется возможность выбора — использовать ли описанный выше блендинг по вероятности или пользоваться моделью

пожара кроны, которая дает наиболее пессимистичный прогноз. Эта модельная скорость распространения пожара может быть разной для разных точек снимка и зависеть от момента времени.

Таким образом, в каждой точке пожара нам известна его скорость распространения. Есть разные подходы к вычислению итоговой максимальной области распространения пожара. Можно составить дифференциальное уравнение и использовать его приближенное решение. Однако мы используем кастомный итеративный алгоритм распространения пожара, действуем отдельно для каждого очага. Строим выпуклую оболочку очага, расширяем ее в соответствии с прогнозом модели распространения пожара. Этот алгоритм позволяет учитывать неоднородность данных как по времени, так и по пространству.

На рис. 5 представлены результаты работы модели: красным и зеленым обозначены разные очаги пожара, контур показывает максимальную область распространения пожара за определенный промежуток времени (в данном случае — 100 часов).

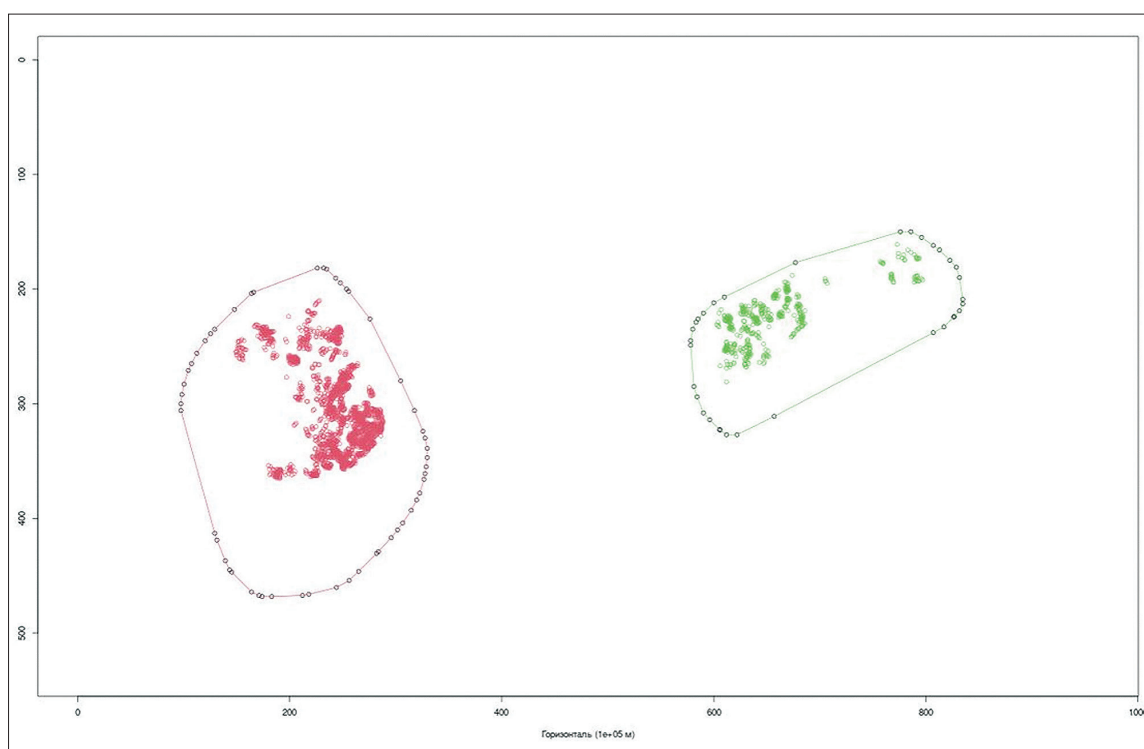


Рис. 5. Результаты работы модели распространения пожаров

Для изображения на рис. 5 использовались исторические данные и учитывалось, что метеорологические данные не однородны как по времени, так и по пространству.

Рис. 7 ниже отображает состояние реки спустя сутки. Видно, что река выглядит чуть больше.

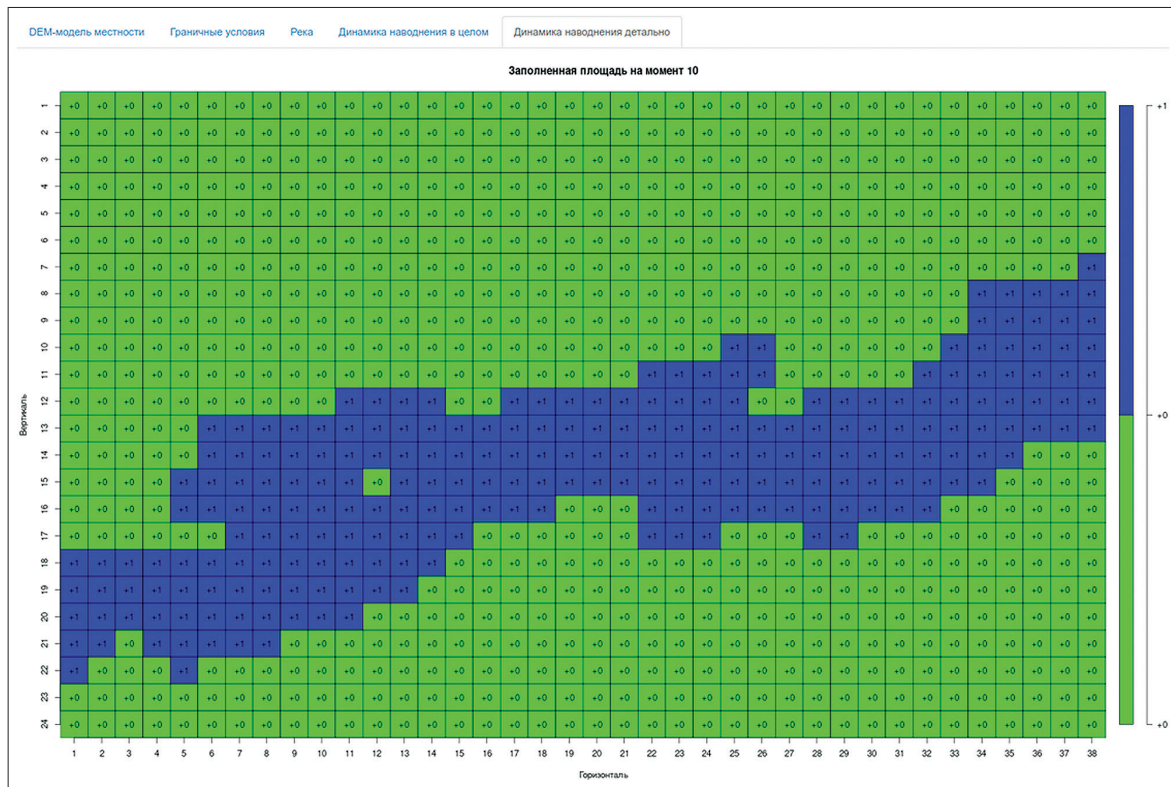


Рис. 7. Река спустя сутки

В качестве основной модели используется модификация модели LISFLOOD-FP из [3]. Так как в некоторых случаях нас не устраивает скорость работы модели, был разработан собственный алгоритм, аппроксимирующий поведение LISFLOOD-FP. Эта идея не нова, в [11] используют сверточные нейронные сети для аппроксимации LISFLOOD-FP.

Основная идея состоит в том, чтобы разбить задачу на две подзадачи: задачу прогнозирования временного ряда количества воды и задачу распространения фиксированного объема жидкости. Первая задача решается с помощью методов классического машинного обучения, методов анализа временных рядов. Вторая задача решается с помощью детерминированного алгоритма, он очень похож на алгоритм, который используется нами в моделировании пожаров легковоспламеняющихся жидкостей и горючих жидкостей.

Начинаем с построения синтетического датасета и применения к нему LISFLOOD-FP. Он делался с помощью малых случайных

возмущений разных имеющихся у нас данных. Его размер составил около 7500 записей.

Далее моделируются относительные приращения количества воды. В качестве метрики берется MAPE, так как она позволяет сравнивать качество моделей временных рядов между различными датасетами. Рассматриваются отдельно интерпретируемые и неинтерпретируемые модели.

При выборе интерпретируемых моделей в силу специфики построения особое внимание уделяется моделям, устойчивым по отношению к выбросам (робастная линейная регрессия). Были опробованы разные модели, однако лучше всего показала себя регрессия Хьюбера [9], была использована ее реализация Huber Regressor в scikit-learn (python). При выборе неинтерпретируемых моделей были рассмотрены самые разные классические варианты. Однако лучше всего показал себя catboost [15], использована его реализация Catboost Regressor в catboost (python). В таблице приведены значения MAPE.

Таблица

Сравнение моделей прогнозирования количества воды

Название модели	MAPE на тренировочном множестве	MAPE на валидационном множестве	MAPE на тестовом множестве
Catboost Regression	2,15%	2,47%	2,39%
Huber Regression	9,23%	8,99%	9,2%

Напомним, что $MAPE < 10\%$ считается очень хорошим.

Несмотря на то, что неинтерпретируемая модель ожидаемо показала себя лучше, чем интерпретируемая, мы рекомендуем к использованию интерпретируемую модель, так как по метрике качества она оказалась достаточно хороша и ожидается, что явная формула будет более устойчивой. Однако, несмотря на эту рекомендацию, в нашем веб-приложении мы оставляем доступными все 3 модели (LISFLOOD-FP, Huberregression, Cat Boostregression), предоставляя пользователю решать, что ему удобнее.

Коротко опишем здесь основную идею алгоритма:

1. Подсчитываются краевые точки реки. Далее будем говорить о соседях краевых точек. Будем считать, что у каждой точки может быть не больше 4 соседей.

2. Действуем итеративно. На каждой итерации есть разлитая вода, и задача распространить еще столько-то воды. Это число может быть как положительным (основной случай), так и отрицательным.

3. Предположим, что стоит задача добавить столько-то воды. Мы итерируемся по краевым точкам и соседям краевых точек, ищем соседа с минимальным объемом и заполняем его. Проверяем, является ли этот сосед краевой точкой, если да, то добавляем его в список краевых точек. Проверяем 4 соседей и пытаемся понять, нет ли среди них кого-то заполненного, кого нужно убрать из списка краевых точек.

4. Предположим, что у нас стоит задача убрать столько-то воды. В этом случае мы действуем в целом аналогично предыдущему пункту.

В ходе имплементации эта идея уточнялась и развивалась.

Заключение

В этой работе описана логика моделирования лесных пожаров и наводнений. Представлены модели детекции пожаров, прогнозирования пожаров, распространения лесных пожаров, распространения наводнений. Подробнее про эти и другие модели можно почитать на сайте ООО «Синкретис» в разделе Исследовательские проекты — Экомоделирование и Smart Geo Platform [18].

Список использованных источников

1. Alexander M.E., Cruz M.G., Evaluating a model for predicting active crown fire rate of spread using wildfire observations, 2005, Canadian Journal of Forest Research 36:2015–3028.
2. Alexandridis A., Russo L., Vakalis D., Bafas G., Wildland fire spread modelling using cellular automata: Evolution in large-scale spatially heterogeneous environments under fire suppression tactics, 2011, International Journal of Wildland Fire, 20(5):633–647.
3. Bates P.D., 2010, The LISFLOOD-FP flood inundation model, University of Bristol.

4. Campello R. J. G. B., Moulavi D., Sander J., Density-Based Clustering Based on Hierarchical Density Estimates, 2013, *Advances in Knowledge Discovery and Data Mining*, 7819:160–172.
5. Farhadi H., Ebadi H., Kiani A., BADI: A novel burned area detection index for Sentinel-2 imagery using Google Earth Engine Platform, 2023, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume X-4/W1–2022.
6. Finney, M.A. 1998. FARSITE: Fire area simulator — model development and evaluation. RMRS-RP-47. USDA ForestServiceRockyMountainResearchStation.
7. Hirsch K.G, Canadian Forest Fire Behavior Prediction (FBP) System: user's guide. 1996. Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Edmonton, Alberta. SpecialReport 7. 122 p.
8. Hendrik H., Schmude J., Predicting Wildfires with Weather Forecast Data, 2021 [Электронныйресурс] URL: <https://developer.ibm.com/exchanges/data/all/spot-challenge-wildfires/> (датаобращения 18.05.2023).
9. Huber P.J., Ronchetti E. M., Robust Statistics Concomitant scale estimates, John Wiley & Sons, Inc.: New Jersey, U.S.A., 2009; p. 172.
10. Ghimire B., Chen A. S., Djordjevic S., Savic D., Application of cellular automata approach for fast flood simulation, 2011, *Computing and Control for the Water Industry*.
11. Kabir S., Patidar S., Xia X., Liang Q., Neal J., Pender G., A deep convolutional neural network model for rapid prediction of fluvial flood inundation, 2020, *Journal of hydrology*, 590.
12. Musabašić M., Fire prediction with logistic regression on territory of Bosnia and Herzegovina, *IOP Conf. Ser.: Mater. Sci. Eng.* 2021.
13. Open-sourceweather API [Электронный ресурс] URL: <https://open-meteo.com/> (дата обращения: 19.05.2023).
14. Rothermel, Richard C. 1991. Predicting behavior and size of crown fires in the northern Rocky Mountains. Res. Pap. INT-RP-438. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 46 p.
15. Prokhorenkova L., Gusev G., Vorobev A., Dorogushh A.V., Gulin A., 2018, Cat-Boost: unbiased boosting with categorical features, 11 pages, Conference NeurIps.
16. Scott, J.H., Reinhardt, E.D. 2001. Assessing crown fire potential by linking models of surface and crown fire behavior. RMRS-RP-29. USDA ForestServiceRockyMountainResearchStation.
17. Van Wagner, C.E. 1993. Prediction of crown fire behavior in two stands of jack pine. *CanadianJournalofForestResearch* 23:442–449.
18. Официальный сайт ООО «Синкретис» [Электронный ресурс] // URL: <https://syncretis.com/ru> (дата обращения: 19.05.2023).