ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ К ЗАДАЧЕ ПРЕДСКАЗАНИЯ ЗЕМЛЕТРЯСЕНИЙ

Галкина А.А., студентка кафедры информационно-аналитических систем СПбГУ, id.a.brickman@gmail.com

Аннотация

Несмотря на чрезвычайную актуальность задачи предсказания землетрясений, на данный момент всё ещё не существует её решения. Это обуславливает необходимость применения современных методов машинного обучения к этой задаче.

В работе описывается создание "эталонного" набора данных, который может быть использован для сравнения различных подходов, предложенных в литературе, а также оценка применимости некоторых методов к данной задаче с использованием этого набора и метрик качества предсказаний.

Введение

Задача предсказания землетрясений критически важна для безопасности человечества. Землетрясения – один из самых страшных и разрушительных природных катастроф, прежде всего, из-за того, что они чаще всего происходят без явного «предупреждения», не оставляя времени на реагирование и принятие мер по смягчению последствий. Помимо этого, землетрясения часто влекут другие бедствия, такие как цунами, лавины, оползни, а также приводят к разрушениям зданий и сооружений и даже могут стать причиной техногенных катастроф (к примеру, авария на АЭС «Фукусима-1» в марте 2011 года произошла вследствие землетрясения, произошедшего в Тихом океане вблизи острова Хонсю и впоследствии признанного сильнейшим в известной истории Японии [1]).

Несмотря на общепризнанную актуальность задачи и постоянные попытки её исследования, берущие своё начало с конца XIX века, на данный момент всё ещё не существует единой методологии, позволяющей строить точные краткосрочные прогнозы сейсмической активности. Это привело к тому, что на данный момент сейсмологи и представители смежных дисциплин скептически относятся к самой возможности построения таких прогнозов [2]. Однако стремительное развитие методов

машинного обучения и успешное применение их к различным классам задач вселяет надежду на то, что эти технологии помогут извлечь скрытые закономерности, на основе которых впоследствии можно будет строить точные предсказания.

Из этих соображений происходит главная цель работы, состоящая в изучении применимости методов машинного обучения к задачам, связанным с предсказанием землетрясений.

Описание задачи

Прежде всего, было необходимо понять, что именно имеется в виду под задачей предсказания землетрясений с точки зрения как сейсмологов, так и представителей смежных дисциплин.

Так, согласно [3], от предсказания землетрясения в его простейшей интерпретации требуется наличие информации о его точном *местоположении*, *промежутке времени*, в который оно случится, а также *диапазоне его магнитуды*. Однако, несмотря на важность проблемы предсказания землетрясений и наличие ясных критериев, которым обязано соответствовать её решение, на данный момент задача построения модели, способной одновременно предсказывать и время, и место, и магнитуду землетрясения, всё ещё является слишком сложной. Поэтому в данный момент исследователи в области машинного обучения чаще всего рассматривают более общую задачу предсказания максимальной магнитуды для некоторого региона и промежутка времени. Под *регионом* при этом подразумевается фиксированный диапазон географических координат (целиком или в виде сеточной матрицы размером, к примеру, 1°×1°). В качестве *временного интервала* чаще всего рассматриваются сравнительно небольшие промежутки времени длиной вплоть до одного месяца.

Что касается классов задач машинного обучения, чаще всего задача предсказания землетрясений сводится к задаче классификации. Это может быть как бинарная классификация (событие произойдёт или не произойдёт) [4-5], так и многоклассовая (события разделяются по диапазонам предсказываемых для них магнитуд) [6-7]. Также встречаются исследования, в которых задача предсказания рассматривается как задача регрессии [8].

Тенденции и проблемы предметной области

Несмотря на столь большое разнообразие рассматриваемых регионов, а также относительную новизну данной сферы исследований, в ходе обзорной части удалось выявить некоторые общие тенденции методологии,

сложившейся в ходе изучения предметной области. Они подробно описаны в опубликованной ранее работе [9]. Ниже изложены те из них, которые повлияли на вектор дальнейших исследований.

Выяснилось, что в ряде работ, где предлагалась некоторая модель и производилась её апробация на нескольких регионах, выявлялось, что один и тот же метод мог подходить для одной сейсмической зоны и не выявить никаких закономерностей в другом случае [4]. Это приводит к существенной проблеме: большое количество исследований использует для создания и апробации моделей данные лишь из одной сейсмической зоны. Помимо этого, для оценки качества предсказаний используются разные (хоть и классические для рассматриваемых задач машинного обучения) метрики. Всё это приводит к тому, что на данный момент всё ещё невозможно объективно позиционировать опубликованные в литературе подходы друг относительно друга.

Возможным решением этой проблемы кажется предоставление исследователям «эталонного» набора данных, на котором можно более полно сравнивать различные алгоритмы машинного обучения – как уже предложенные, так и те, которые будут разработаны в будущем. И разработка такого инструмента стала первой задачей практической части данной работы.

«Эталонный» набор данных

Итак, в ходе работы был создан так называемый «эталонный» набор данных, содержащий данные о землетрясениях из различных сейсмических зон (их описание приведено в Табл.1). Выбор регионов обусловлен тем, что некоторые из них (Южная Калифорния, Чили, Хиндукуш) уже неоднократно рассматривались в опубликованных работах. Другие же выбраны из соображений репрезентации различныхчастей света в результирующем наборе. В силу того, что различные источники данных в разной степени полны для разных регионов, было принято решение использовать не одну базу данных о землетрясениях, а несколько разных. Также была произведена первичная обработка данных, а именно, обеспечение их полноты путём выявления отсекающей (*cut-off*) магнитуды – границы, начиная с которой набор можно считать полным [9].

Помимо «сырых» исторических данных, в набор было решено включить подсчитанные значения описанных в литературе*сейсмических индикаторов*–параметров, основанных на сейсмологических соотношениях и гипотезах, обобщающих динамику сейсмичности в регионах [4][6]. Это позволит исследователям предметной области обучать и апробировать модели без необходимости разбираться в сложных

сейсмологических законах.

Результирующий набор данных, а также ряд полезных функций для предобработки данных и извлечения сейсмических индикаторов доступен по ссылке: https://github.com/abrickman/benchmark-earthquake-dataset.

Сейсмические зоны	**Диапазоны широт**	**Диапазоны долгот**	**Отсекающая магнитуда**
Центральная Япония	34°-39° с.ш.	136.5°-142°в.д.	4.5
Центральное Чили	32.5°-36° ю.ш.	70°-72.5° з.д.	4.0
Хиндукуш, Пакистан	35°-39° с.ш.	69°-74.6° в.д.	4.5
Сицилия, Италия	36°-39° ю.ш.	12°-16° в.д.	2.5
Южная Калифорния	32°-36.5° с.ш.	114.75°-121° з.д.	3.0

Таблица 1. Описание «эталонного» набора данных о землетрясениях

Эксперименты

Важной задачей работы стала апробация методов машинного обучения на собранных данных с целью проверки их применимости к задаче предсказания на разных регионах. Задача формулировалась как задача предсказания максимальной магнитуды в регионе за 7 дней и рассматривалась с точки зрения бинарной классификации: 1 или 0 в зависимости от того, случилось ли в регионе землетрясение с магнитудой, превысившей пороговое значение. При этом пороговая магнитуда выбиралась для каждого региона в отдельности как значение «выше среднего». Что касается методов, в работе были рассмотрены такие алгоритмы классификации, как *логистическая регрессия*, классификаторы на основе *k ближайших соседей*, *метод опорных векторов*, *деревья решений* и *случайные леса*. Для них был осуществлён подбор оптимальных параметров методом поиска по сетке. Также в рамках экспериментов была осуществлена попытка поиска оптимальной нейросетевой архитектуры (среди имеющих топологию прямого распространения трёх- и четырёхслойных сетей с сигмоидальной функцией активации).

Метрики качества предсказаний

Для оценки качества рассматриваемых моделей в работе использовались метрики, основанные на элементах т.н. матрицы ошибок (*confusionmatrix*), отражающей все возможные исходы:

	Произошло	**Не произошло**
Было предсказано	$\boldsymbol{TP}$	FP
Не было предсказано	FN	$\boldsymbol{TN}$

Таблица 2. Матрица ошибок моделей для предсказания землетрясений

Ниже даны определения метрик качества, использованных в данном исследовании. Прежде всего, это точность (*accuracy*), указывающая на долю правильных предсказаний среди всех предсказаний, сделанных моделью:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}.$$

Также модели оценивались с точки зрения таких критериев, как чувствительность (*sensitivity*, обозн. S_n илиPOD - от*probability of detection*) испецифичность (*specificity*, обозн. S_p):

$$S_n = \frac{TP}{TP + FN}; \qquad S_p = \frac{TN}{TN + FP}.$$

Наконец, важными критериями оценки качества выступают метрики P_1 (*positivepredictivevalue*) и P_0 (*negativepredictivevalue*), которые определяются следующим образом:

$$P_1 = \frac{TP}{TP + FP}; \qquad P_0 = \frac{TN}{TN + FN}.$$

Выводы на основе результатов

Итак, в ходе экспериментов, прежде всего, было установлено, что методы машинного обучения могут быть использованы для решения задач, связанных с предсказанием землетрясений. При этом, как показывают значения метрик для моделей, построенных по данным для Центрального Чили (см. Табл. 3), эти методы могут давать хорошие результаты как с точки зрения чувствительности к землетрясениям, так и в смысле небольшого количества «ложных тревог», что также является существенным.

	Logistic Regression	*KNN Classifier*	*Support Vector Classifier*	*Decision Tree*	*Random Forest*	*Feed-forward ANN*
$\boldsymbol{P_1}$	0.62	0.54	0.50	0.52	0.52	**0.62**
$\boldsymbol{P_0}$	0.61	0.76	0.79	0.67	0.70	**0.72**
$\boldsymbol{S_n}$	0.13	0.75	0.84	0.51	0.62	**0.57**
$\boldsymbol{S_p}$	0.95	0.56	0.43	0.67	0.60	**0.76**
Accuracy	0.61	0.64	0.59	0.60	0.61	**0.68**

Таблица 3. Результаты работы моделей на наборе индикаторов Reyes et al. [4] для Центрального Чили

Однако в процессе построения и оценки качества моделей было выявлено сильное различие в результатах, полученных для разных

регионов. С точки зрения сбалансированности, чувствительности и точности модели (в том числе нейросетевые, как показано в Табл. 4), построенные для данных из некоторых зон оказались в среднем хуже, чем классификаторы, обученные на данных из Чили.

Feed-forward ANN	*Chile, AP [6]*	*Chile, Reyes [4]*	*Sicily, AP [6]*	*Sicily, Reyes [4]*	*South California, AP [6]*	*South California, Reyes [4]*
$\boldsymbol{P_1}$	0.55	0.62	0.44	0.47	0.23	0.50
$\boldsymbol{P_0}$	0.68	0.72	0.59	0.60	0.81	0.81
$\boldsymbol{S_n}$	**0.52**	**0.57**	**0.14**	**0.15**	**0.18**	**0.24**
$\boldsymbol{S_p}$	0.71	0.76	0.87	0.88	0.86	0.99
Accuracy	0.63	0.68	0.57	0.58	0.73	0.80

Таблица 4. Результаты работы нейронных сетей прямого распространения на трёх сейсмических зонах (Чили, Сицилия, Южная Калифорния)

Мало того, даже в пределах одного региона для разных наборов сейсмических индикаторов получались модели разного качества. Примером такого региона стала Центральная Япония: в случае использования набора индикаторов Adeli & Panakkat [6] результирующая нейронная сеть прямого распространения гораздо лучше улавливает землетрясения, чем построенная на основе индикаторов Reyes et al. [4], а классификатор на основе k ближайших соседей показывает наиболее сбалансированные результаты в смысле рассматриваемых метрик.

Adeli & Panakkat [6]	*KNN Classifier*	*Feed-forward ANN*
$\boldsymbol{P_1}$	0.52	0.49
$\boldsymbol{P_0}$	0.70	0.62
$\boldsymbol{S_n}$	**0.69**	**0.44**
$\boldsymbol{S_p}$	0.52	0.67
Accuracy	0.60	0.57

Reyes et al. [4]	*KNN Classifier*	*Feed-forward ANN*
$\boldsymbol{P_1}$	0.46	0.59
$\boldsymbol{P_0}$	0.59	0.58
$\boldsymbol{S_n}$	**0.42**	**0.05**
$\boldsymbol{S_p}$	0.64	0.97
Accuracy	0.55	0.58

Таблица 5. Сравнение работы моделей, построенных на основе разных сейсмических индикаторов (Центральная Япония)

Подобный результат можно объяснить тем, что сейсмичность в разных зонах имеет разную природу, и наборы сейсмических индикаторов, соответствующие различным моделям распределения сейсмичности, могут вносить различный «вклад» в активность в том или ином регионе. Соответственно, для того, чтобы уметь предсказывать землетрясения в регионе, необходимо понимать процессы, влияющие на его сейсмические свойства, а значит, требуется индивидуальный подход к подбору наиболее релевантных индикаторов для рассматриваемых сейсмических зон.

Заключение

Итак, суммируем результаты, достигнутые на данный момент: во-первых, в ходе обзора последних исследований [9] удалось выявить общие черты методологии, а также установить основные проблемы, тормозящие развитие предметной области. Решению одной из них была посвящена практическая часть, в ходе которой был создан и размещён в открытый доступ «эталонный» набор данных о землетрясениях, обогащённый сейсмическими индикаторами. Наконец, на полученных данных был проведён ряд экспериментов, продемонстрировавших потенциал применения методов машинного обучения к данной задаче. Полученные результаты не являются финальными и абсолютными, однако они могут стать как отправной точкой для других исследователей в этой области, так и фундаментом для моих дальнейших исследований, перспективные направления которых также выделены в рамках работы.

Литература

1. N. Mimura, K. Yasuhara, S. Kawagoe, H. Yokoki, and S. Kazama, “Damage from the Great East Japan Earthquake and Tsunami - A quick report,” Mitigation and Adaptation Strategies for Global Change, vol. 16(7), pp. 803–818, 2011.
2. R. Geller, “Earthquake prediction: a critical review,” Geophysical Journal International, vol. 131, pp. 425–450, 2007.
3. D. D. Jackson, “Hypothesis testing and earthquake prediction,” Proc. Natl. Aca. Sci. USA, vol. 93, pp. 3772–3775, 1996.
4. F. Martínez-Álvarez, A. Morales-Esteban, and J. Reyes, “Neural networks to predict earthquakes in Chile,” Applied Soft Computing, vol. 13, pp. 1314-1328, 2013.
5. K. Asim, A. Idris, T. Iqbal, and F. Martínez-Álvarez, “Seismic indicators based earthquake predictor system using Genetic Programming and AdaBoost classification,” Soil Dynamics and Earthquake Engineering, vol. 111, pp. 1–7, 2018.
6. A. Panakkat, and H. Adeli, “Neural network models for earthquake magnitude prediction using multiple seismicity indicators,” International Journal of Neural Systems, vol. 17(1), pp. 13–33, 2007.
7. H. Adeli, and A. Panakkat, “A probabilistic neural network for earthquake magnitude prediction,” Neural Networks, vol. 22(7), pp. 1018–1024, 2009.
8. G. Cortés, A. Morales-Esteban, X. Shang, and F. Martínez-Álvarez, “Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure,” Computers & Geosciences, vol. 115, pp. 198-210, 2018.

9. A. Galkina, and N. Grafeeva, "Machine learning methods for earthquake prediction: a survey", CEUR Workshop Proceedings, vol. 2372, pp.25-32, 2019.