Высшая школа бизнеса

Департамент бизнес-информатики

Москва
2023

Оценка потребностей малого бизнеса сейсмоопасных районов в ресурсах

Научный руководить: доктор технических наук, заслуженный профессор,
Алескеров Фуад Тагиевич,
Студент: Сторожок Мария Константиновна, группа 1908

Содержание

Цель, задачи, объект и предмет исследования

Цель работы

Повысить качество получаемых оценок потребностей в ресурсах, необходимых для устранения последствий сейсмических событий, путем расширения исследований применимости различных методологий и типов исходных данных для **прогнозирования сильных землетрясений**.

Задачи

1. Разработка нового метода, базирующегося на применении случайного леса к геофизическим и сейсмологическим данным,
2. Создание набора данных,
3. Изучение влияния комбинирования двух типов данных на качество модели.

Объект и предмет

Объектом является **более точное предсказание землетрясений**, что позволит повысить надежность прогнозов потребностей в ресурсах. Предметом является применимость **метода случайного леса и двух выбранных типов данных** для построения более точных прогнозов сильных землетрясений.

Актуальность для бизнес-заказчика

Сильные землетрясения магнитудой 8 и выше случаются примерно **раз в год**. Вследствие этого в сейсмоопасных районах по всему миру ежегодно из-за не предсказанных землетрясений останавливаются производства, страдают торговые и коммерческие помещения многих компаний, **бизнесы теряют прибыль**.

Кейс Японии:

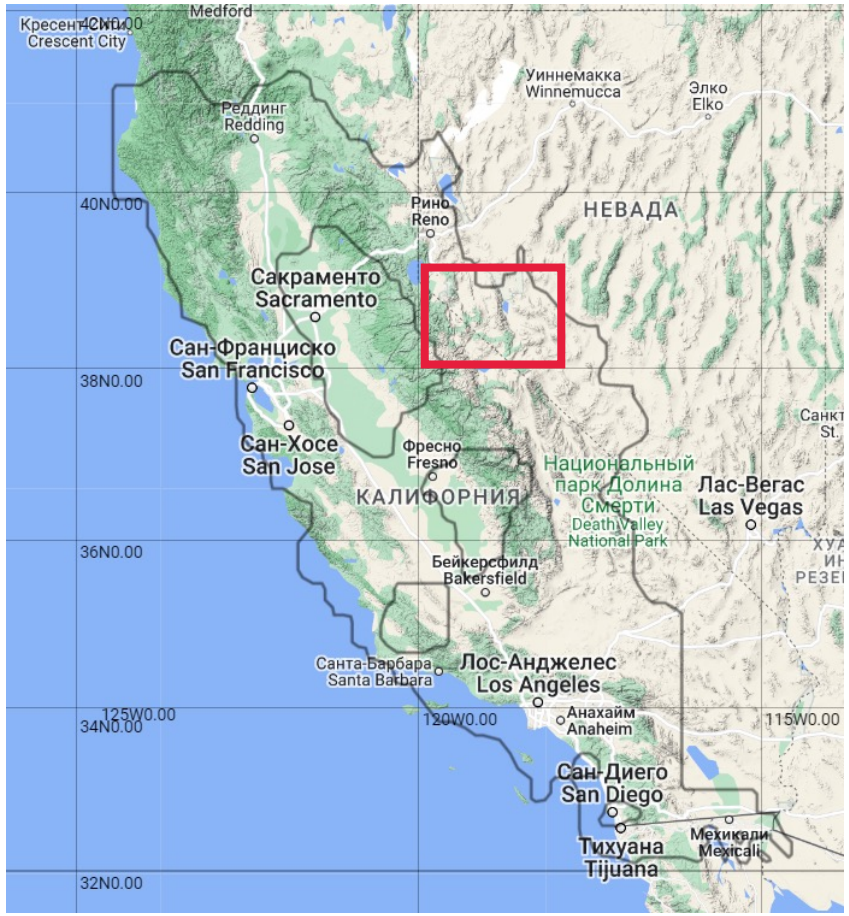
Ущерб от землетрясения в Японии 11 марта 2011г. оценивается в **309 млрд долл**. **Прибыль крупнейшего в мире автоконцерна Toyota Motor упала в тот год на 99,4%** - до 1,6 млрд иен (15 млн долл.). Это почти в 165 раз меньше, чем показатель за тот же квартал прошлого года, когда Toyota заработала 190,5 млрд иен (2,4 млрд долл.). Другие японские автопроизводители также понесли существенный потери. Так**, прибыль Honda упала в 9 раз, Nissan - на 20%, а убытки Mazda увеличились в 12 раз**. И все это из-за одного не предсказанного землетрясения!

Такие **убытки можно было бы избежать**, если бы сейсмический удар был **вовремя предсказан**.

Источник: https://www.iris.edu/hq/inclass/fact-sheet/how_often_do_earthquakes_occur, https://www.rbc.ru/society/02/08/2011/5703ea349a79477633d3607e

Входные данные: область прогнозирования

- Территория, расположенная в границах 119,5–117,4° з.д. и 38–39° с.ш. в Калифорнии, делится на отдельные участки размером 0,1° × 0,075°
- Для каждого участка раз в 30 дней вычисляются геофизические и сейсмологические параметры, аномальные значения которых могут быть предвестниками сильных землетрясений
- В работе рассматривается период с 7 августа 2009 по 26 января 2023 года
- Итого: 315 участков, 165 временных шагов и 43 413 записей (строк) в датасете после удаления пропусков
- Значения признаков нормализуются по методу минимакса



Источник: Gitis V., Derendyaev A., Petrov K. Analyzing the performance of GPS data for earthquake prediction // Remote Sensing. – 2021. – Vol. 13, № 9.

Входные данные: предсказательные переменные

Геофизические признаки

Временные ряды координат приемных станций GPS по суточным данным

Суточные горизонтальные смещения земной поверхности в местах расположения станций и на целых участках

Дивергенция (расходимость), ротор и сдвиг скоростей деформации

Изменения инвариантов скоростей деформации во времени

Сейсмологические признаки

Значение плотности эпицентров землетрясений и его квантиль

t-статистика Стьюдента средней магнитуды землетрясений

Источник: Gitis V., Derendyaev A., Petrov K. Analyzing the performance of GPS data for earthquake prediction // Remote Sensing. – 2021. – Vol. 13, № 9.

Входные данные: целевая переменная

- Из архива землетрясений, произошедших за рассматриваемый период и имевших магнитуду 5 и более, удаляются записи о форшоках и афтершоках
- Остается 8 землетрясений
- Каждой записи в датасете ставится в соответствие 1, если по расположенным в ней значениям параметров необходимо предсказать землетрясение в следующие 30 дней, иначе – 0
- Предсказание магнитуды затруднительно ввиду малого числа землетрясений, но так как большая часть землетрясений в Калифорнии имеют магнитуду 5-6,5 и более сильные толчки происходят крайне редко, получив предсказание 1, можно с большой долей вероятности ожидать событие магнитудой не более 6-6,5, что позволит эффективно предсказывать потребность в ресурсах, а значит, такого формата предсказаний в данном регионе может быть вполне достаточно

Источник: Gitis V., Derendyaev A., Petrov K. Analyzing the performance of GPS data for earthquake prediction // Remote Sensing. – 2021. – Vol. 13, № 9.

Входные данные: борьба с дисбалансом классов

- 40 413 записей принадлежат классу 0 и только 8 записей – классу 1
- Даже после замены задачи регрессии на задачу классификации модель не сможет делать качественные прогнозы
- Генерируется 20 новых объектов класса 1, значения признаков для которых случайным образом выбираются из значений реальных объектов того же класса из набора данных

Пример строк из итогового набора данных

Год	Месяц	День	Долгота	Широта	F4	F5	F6	S9	S11	Класс
2009	8	7	-119,5	39,007	0,503	-0,219	1,245	0,446	2,142	0
2009	8	7	-119,4	39,007	2,261	3,551	8,822	0,523	2,231	0
2009	8	7	-119,3	39,007	-8,884	8,894	-8,575	0,524	2,312	0
2009	8	7	-119,2	39,007	-20,893	-0,0514	-15,676	0,654	2,377	0
2009	8	7	-118,5	39,007	14,297	-10,377	-15,117	0,249	2,5127	0

Обучение и тестирование модели

- Объекты класса 1 разбиваются в соотношении 70:30
- Обучаются 100 решающих деревьев со следующими гиперпараметрами: максимальная глубина дерева = 2, минимальное число семплов в листе = 2
- Каждое на случайной подвыборке объектов класса 0 и фиксированных 70% объектов класса 1
- Объектов класса 0 берется столько же, сколько элементов класса 1 попало в обучающую выборку
- Аналогично создается тестовая выборка

Матрица путаницы

Предсказанные классы \ Реальные классы	1	0
1	True Positive (TP)	False Positive (FP)
0	False Negative (FN)	True Negative (TN)

- Значения метрик считаются для каждого дерева и усредняются
- Чем лучше в среднем значения метрик деревьев, тем лучше в среднем предсказания каждого из деревьев, а значит, надежнее итоговый прогноз случайного леса

$$POD = \frac{TP}{TP + FN}, \quad FAR = \frac{FP}{FP + TN}, \quad R = POD - FAR.$$

Результаты применения модели

Геодинамические признаки	**Сейсмологические признаки**	**Два типа признаков**
POD = 0,751,	POD = 0,462,	POD = 0,762,
FAR = 0,184,	FAR = 0,04,	FAR = 0,135,
R = 0,567	R = 0,422	**R = 0,627**

Преимущества и ограничения модели

Преимущества модели

Прогноз строится для небольшого участка, что позволит эффективно оценивать потребность в ресурсах и оперативно принимать необходимые меры

Ограничения модели

- Использование ограниченного набора данных
- Использование данных из одного региона
- Подмена целевой переменной

Рекомендации по улучшению модели

- Использование набора данных большего размера за счет расширения области анализа
- Рассмотрение иных способов синтеза и аугментации данных
- Переход от решения задачи двухклассовой классификации к предсказанию более сложных целевых переменных, например, уровня опасности
- Использование модели улучшенного дерева решений, предложенной Алескеровым Ф. Т. и др. в 2020 году

Источник: Aleskerov F., Baiborodov N., Demin S., Shvydun S., Trafalis T., Richman M., Yakuba V. Constructing an efficient machine learning model for tornado prediction // International Journal of Information Technology & Decision Making. – 2020. – Vol. 19, № 5. – P. 1177–1187.

Заключение

1 модель

Метод случайного леса

2 типа данных

Геофизические и сейсмологические данные

1 регион

Область в восточной части Калифорнии

Подведение итогов

Какие были получены результаты работы?	**П**олучение прогнозов сейсмических ударов с помощью предложенного метода позволит бизнесам штата Калифорния **более точно оценивать свои потребности в ресурсах**, а также **предоставлять поддержку** своим сотрудникам и клиентам, государству и другим жителям региона.
Где в работе бизнес?	**С**нижение риска получения недостаточной прибыли из-за неопределенности для бизнеса в сейсмоопасном районе.
Где в работе информатика?	**И**спользование подхода, благодаря которому компьютер может анализировать данные и обучаться на их основе.
Какие были проведены исследования?	**И**сследованы изученные ранее подходы к прогнозированию землетрясений.
Что было сделано самостоятельно?	**С**качены и обработаны данные, собран датасет, обучена и протестирована модель.

Список основных источников

Gitis V., Derendyaev A., Petrov K. Analyzing the performance of GPS data for earthquake prediction // Remote Sensing. – 2021. – Vol. 13, № 9.

Mignan A., Broccardo M. Neural network applications in earthquake prediction (1994–2019): Meta-analytic and statistical insights on their limitations // Seismological Research Letters. – 2020. – Vol. 91, № 4. – 2330–2342.

Aleskerov F., Say A. I., Toker A., Akin H. L., Altay G. A cluster-based decision support system for estimating earthquake damage and casualties // Disasters. – 2005. – Vol. 29, № 3. – P. 255–276.

Aleskerov F., Baiborodov N., Demin S., Shvydun S., Trafalis T., Richman M., Yakuba V. Constructing an efficient machine learning model for tornado prediction // International Journal of Information Technology & Decision Making. – 2020. – Vol. 19, № 5. – P. 1177–1187.

Вопросы

Спасибо за внимание!

Остались ли у Вас вопросы?

Геофизические параметры

Исходные данные представляют собой временные ряды суточных координат x(t) и y(t) приемных станций GPS, расположенных в исследуемой области, в направлениях З-В и С-Ю. По двум координатам станции GPS, зарегистрированным с временным интервалом T_0 дней, определяются суточные горизонтальные смещения земной поверхности $g_x(t)$ и $g_y(t)$:

$$g_x(t) = \frac{x(t) - x(t - T_0)}{T_0}.$$

Скорость деформации земной поверхности в направлении С-Ю $g_y(t)$ рассчитывается аналогично.

Для расчета суточных скоростей используется интервал T_0 = 30 дней.

Геофизические параметры

Далее вычисляются скорости смещения V_x и V_y в направлениях З-В и С-Ю для каждого участка:

$$V_{xn}(t) = \frac{\sum_{k=1}^{K} g_{xk}(t) / r_k^p}{\sum_{k=1}^{K} 1 / r_k^p},$$

где К – максимальное число ближайших к участку n станций в радиусе R_{max}, значения которых использовались для вычисления скорости деформации,

$r_k <= R_{max}$ – расстояние от k-ой станции до участка n,

p – степень, определяющая зависимость веса станции от ее расстояния до участка.

В настоящем исследовании использовались следующие значения параметров: К = 5, R_{max} = 50 км, p = 1. Соответствующие расчеты скорости деформации в направлении С-Ю проводятся аналогичным образом.

Геофизические параметры

- Аномальные значения параметра F_1, или дивергенции (расходимости) скоростей деформации, соответствуют участкам, где происходит относительное сужение или расширение размера небольшой горизонтальной поверхности :

$$divV_n = \frac{\partial V_{xn}}{\partial x} + \frac{\partial V_{yn}}{\partial y}.$$

- F_2, или ротор скоростей деформации, определяет направление и интенсивность закручивания участка вокруг вертикальной оси:

$$rotV_n = \frac{\partial V_{xn}}{\partial y} - \frac{\partial V_{yn}}{\partial x}.$$

- F_3 определяет сдвиг скоростей деформации

$$shV_n = \frac{1}{2}\sqrt{\left(\frac{\partial V_{xn}}{\partial x} - \frac{\partial V_{yn}}{\partial y}\right)^2 + \left(\frac{\partial V_{xn}}{\partial y} + \frac{\partial V_{yn}}{\partial x}\right)^2}.$$

Геофизические параметры

Параметры F_4, F_5, F_6 показывают изменения инвариантов скоростей деформации F_1, F_2, F_3 во времени. Каждый из них равен отношению разности средних значений определенного инварианта в двух последовательных временных интервалах к стандартному отклонению этой разности и вычисляется раз в 30 дней.

F_4 – признак временных вариаций дивергенции скоростей деформации:

$$f_{4n}(t) = \frac{\overline{div_{2n}} - \overline{div_{1n}}}{\sigma_n(div)},$$

где div_{2n} рассчитывается по значениям параметра F_1 на интервале $(t - T_2, t)$, div_{1n} – по значениям параметра F_1 на интервале $(t - T_2 - T_1, t - T_2)$, $T_1 = T_2 = 360$ дней.

Признаки F_5 и F_6 вычисляются аналогично.

Сейсмологические параметры

- S_1, или плотность эпицентров землетрясений, вычисляется методом локальной ядерной регрессии. Функция ядра для q-го землетрясения имеет вид:

$$K_q = \left[\cosh^2\left(\frac{r_q}{R}\right)^2 \cosh^2\left(\frac{t_q}{T}\right)\right]^{-1},$$

где $r_q < R$ – расстояние между эпицентром q-го землетрясения и участком,

$t_q < T$ – временной интервал между эпицентром q-го землетрясения и участком.

Используются следующие значения параметров: R = 50 км, T = 100 дней.

- S_8 равняется квантилю плотности эпицентров землетрясений, и на его основе вычисляется параметр S_9:

$$S_9 = \frac{S_1}{S_8 + 0{,}001}.$$

Сейсмологические параметры

- Параметр S_{11} основывается на t-статистике Стьюдента средней магнитуды землетрясений и определяется для каждого участка как отношение разности средних значений текущего интервала T_2 и фонового интервала T_1 к стандартному отклонению этой разности. Используются временные промежутки $T_1 = 2310$ и $T_2 = 120$ дней.