Оценки риска возникновения лесных пожаров методами машинного обучения

A Preprint

Zharov Georgii
MIPT
zharov.g@phystech.edu

Юрий Максимов

Abstract

В работе рассматривается проблема прогнозирования эксктримальных климатических явлений, а именно лесных пожаров. В качестве цели ставиться предсказание пожаров на основе уже имеющихся данных о явлениях подобного типа методами глубокого обучения. Для краткосрочного прогнозирование исследуются стационарные временные ряды. Для осуществствления долгосрочного прогнозирования исследуются нестационарные временные ряды.

Keywords First keyword · Second keyword · More

1 Introduction

Прогнозирование экстремальных климатических явлений является важной прикладной задачей, так как природные катаклизмы могут нанести существенный вред многим сферам человеческого хозяйства. Целью данной работы является постороения модели для предсказания появлений лесных пожаров. Для решения поставленной задачи предлагается исследовать поведение временных рядов, стационарных - для прогнозирования явлений в течение временного промежутка порядка $4 - 5$ лет, и нестационарных - для предсказания в явлений в теничение $40 - 50$ лет.

Основная сложность в решении задачи прогнозировования экстримальных явлений заключается в том, что эти явления являются случайными и происходят достаточно редко - интервал времени между двумя соседними событиями в рассматриваемом регионе может быть весьма большим. В силу этой особенности исследуеиого явления, при решении задачи мы можем столкнуться с проблемой несбалансированных данных. Эта проблема заключается в следущем. При работе с данными мы вводим определенные пороговые значения для рассмматриваемых признаков, и события, для которых значения признаков превосходят порговые, мы будем считать экстремальными. Так как интересующие нас события происходят достаточно редко, лишь малая часть событий в выборке будут помечены как как экстримальные. Таким образом, большая часть данных будет находится в пределах пороговых значений, а наиболее интересные для нас события будут составлять меньшинство, и выборка будет несбалансированной. Из-за несбалансированности входных данных при обучении модели мы можем столкнуться как с проблемой недообучения, так и с проблемой переобучения, примеры этого приведены в работе [1].

Основной целью работы является поиск архитектуры, которая будет лучше справлятся с задачей прогнозирования поведения временных рядов и предсказанием экстремальных событий чем уже существующие модели. Классические модели плохо справляются с поставленной задачей. Поэтому для решения предлается использовать архитектуру с элементами рекуррентных нейронных сетей (RNN), способных хранить в памяти информацию о предшествующих событиях. Также в работе [1] можно найти объснение почему использование квадратиной функии потерь в данной задаче приводит к плохим

результатам. Это происходит из-за того большинство часто используемых распределений, например распределение Гаусса или Пуассона, не описывают heavy-tailed данные. Так называются данные, среди которых в малом количестве содержатся события с маленькой веротностью происхождения. В нашей задаче экстремальные события можно интепретировать как такие данные. Для решения этой предлагается добавить к квадратичной функции потерь так называемую Extreme Value Loss (EVL) [1]. Такое решение позволяет учитывать heavy-tailed данные и корректно обучать на них модель.

В качестве основного датасета исползуется информация о лесных пожарах на территории США за предшедтвующие несколько десятков лет с сервиса Google Earth Data. Также используются данные из Wildfire Risk Database и Severe Weather Dataset.

2 Problem statement

В качестве входных данных выступают набор двумерных географических точек (x_i^1, x_i^2), время t и вектор климатических параметров h. Предсказание нашей модели для момента времени o_t. Выходные данные для момента времени t это y_t. $T-$ количество рассматриваемых временных точек. Мы можем взять квадратичную функцию потерь и поставить задачу минимизации следующим образом

$$\min \sum_{t=1}^{T} \|o_t - y_t\|^2$$

Но, как было сказано во введении, такой подход не будет оптимальным. Вместо этого, в соответствии с алгоритмом предложенным в статье [1], мы можем выбрать функцию потерь и поставить задачу оптимизации так

$$\min \sum_{t=1}^{T} (\|o_t - y_t\|^2 + \lambda_1 EVL(w_t, v_t))$$

где $v_t = \{0, 1\}$ это индикатор, показывающий, считаем ли мы данное собыытие экстремальным или нет, λ_1 и w_t это параметры.

Список литературы