Построение предиктивной аналитики для сенсоров загрязнений атмосферы

A Preprint

Кристина М. Казистова*
ФПМИ
МФТИ
Долгопрудный
kazistova.km@phystech.edu

Abstract

Загрязнение воздуха - это явление, вредное как для существования человека, так и для экологической системы. Оно вызвано избытком некоторых веществ в атмосфере выше определенной концентрации. По этой причине актуальной проблемой является изучение новых методов прогнозирования будущего загрязнения воздуха. В данной работе мы будем предсказывать концентрацию загрязняющих веществ в воздухе на неделю вперед с интервалом в 5 минут. Будут учтены метеорологические факторы, выбросы загрязняющих веществ в режиме реального времени, расположение сенсоров.

1 Introduction

Повсеместное ухудшение качества воздуха как в развивающихся, так и в развитых странах привело к возникновению глобальной угрозы, оказывающей огромное негативное воздействие на окружающую среду и здоровье человека. Загрязнение воздуха рассматривается как смесь газов и частиц, которая концентрируется во вредных количествах, выбрасываемых в атмосферу. Одними из основных источников, вызывающих загрязнение воздуха, являются транспортные выбросы и сжигание топлива. Последствия загрязнения воздуха проявляются в виде двух основных воздействий: воздействия на здоровье человека, такого как болезни, и воздействия на окружающую среду, такого как глобальное потепление, изменение климата, кислотные дожди и загрязнение твердыми частицами.

В силу вышеперечисленных фактов многие исследовательские усилия направлены на прогнозирование концентрации загрязняющих веществ в воздухе в будущем. Такие прогнозы позволяют контролировать качество воздуха, разрабатывать стратегии борьбы с его загрязнителями и своевременно принимать меры предосторожности. Данная задача очень актуальна и еще недостаточно хорошо изучена, что открывает простор для исследований.

Нам предстоит строить предсказания сенсоров загрязнений атмосферы в Москве и Московской области. Сенсоры показывают концентрацию различных веществ, содержащихся в воздухе, через фиксированные промежутки времени. Показания сенсоров представляют собой временной ряд - значения меняющихся во времени признаков, полученных в некоторые моменты времени. Помимо значений концентрации вредных веществ мы будем учитывать метеорологические данные, такие как скорость и направление ветра, температура, давление и другие, а также местоположение сенсоров относительно друг друга. В итоге мы получим многомерные временные ряды, которые очевидным образом скоррелированы между собой (близко расположенные станции дают похожие показания). Получаем, что предложенный метод решения должен учитывать как пространственные, так и временные зависимости.

*Use footnote for providing further information about author (webpage, alternative address)—not for acknowledging funding agencies.

1. Здесь не дефис, а длинное тире (в латехе пишется как три дефиса, ---)

2. Если статья на русском, лучше писать обезличенно:
мы будем предсказывать => предлагается предсказывать

3. И лучше писать все изложение в настоящем времени

4. Не хватает информации, на каких данных проводится эксперимент

5. Хотелось бы какую-нибудь ссылку на работу, где говорится об этих вещах

6. Опять же, лучше писать более уверенно и обезличенно.
Вариант: рассматривается задача построения модели предсказания сенсоров...

Существующие исследования прогнозирования в целом можно разделить на три категории, основанные на методах моделирования, а именно: численные методы, статистические методы и методы машинного обучения.

Численные модели широко внедряются в системы прогнозирования, однако для достижения высококачественных результатов входные данные требуют большей точности, чем та, которая доступна в настоящее время. Кроме того, большинство практических ситуаций с воздушной средой сложны и их трудно выразить математически.

Статистические методы обосновываются путем проверки правильности гипотезы распределения вероятностей для данных. Однако предположения о распределении вероятностей данных также подразумевают ограничения для дальнейших практических приложений.

Наконец, все более распространенными становятся методы прогнозирования, основанные на технологиях машинного обучения. Такого рода модели непосредственно исследуют сложные скрытые закономерности в данных, не требуя ни гипотетического распределения переменных или данных, ни глубокого понимания физических или химических свойств загрязнителей воздуха.

Изучение литературы показало, что хорошим решением является использование модели LSTM, которая является популярным вариантом метода рекуррентной нейронной сети (RNN). Данная модель учитывает влияние предыдущих значений на текущее при расчете модели. Эта особенность делает LSTM одной из наиболее подходящих моделей для задач прогнозирования качества воздуха, поскольку временная зависимость является типичным явлением, наблюдаемым в рядах концентраций загрязнителей воздуха.

2 Problem statement

Пусть N - это количество датчиков в интересуемой нас области. Сопоставим каждому датчику с номером i временной ряд x_i - показания этого датчика с течением времени. Введем следующие обозначения: d - количество различных показателей, которые регистрирует датчик (различные загрязняющие атмосферу вещества), L_k - текущее количество равных временных промежутков, для которых нужно сделать предсказания, U_j - дополнительные факторы, такие как метеорологические условия, расстояния между станциями и прочее.

Пусть f_k - функция из класса нейронных сетей, предсказывающая значения датчиков в следующие L_k промежутков времени по показаниям за предыдущие L_{k-1} промежутков, учитывающая дополнительные факторы. Тогда

$$f(x^{N,L_{k-1},d}, U_{L_{k-1}}, w) = \hat{x}^{N,L_k,d},$$

где w - параметры модели, обучаемые путем минимизации функции потерь.

8. Текст получается очень длинный. Сократить раза в полтора. Нужны ссылки по каждой группе методов

9. Не хватает абзац с описанием что будете делать в работе вы. (этот абзац нужен, даже если потом планы поменяются)

10. Постановка задачи

11. Есть стандартная нотация - векторы пишут жирным, матрицы - жирным и заглавными буквами

12. Какие факторы, к какому множеству принадлежат?

13. Какая функция потерь? Расписать