Применение алгоритмов машинного обучения для заполнения пропущенных значений в данных о загрязнении почвы

Лебедева Мария Александровна

Студент, магистр

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: mash.lebedeva2010@yandex.ru

Научный руководитель *— Королев Виктор Юрьевич*

Анализ загрязнения почвы, важного параметра состояния окружающей среды, на практике осложняется не только природой своей изменчивости, но и наличием пропущенных значений, которые препятствуют применению статистических методов работы с временными рядами. Алгоритмы обработки данных часто очень чувствительны к наличию пропусков, поскольку они могут серьезно исказить результаты анализа. В частности, статистические тесты и пороги для экстремальных значений загрязнения, основанные на неполных данных, могут быть неверными [1]. Следовательно, требуется эффективный метод для определения количества загрязняющих веществ в почве. В данном исследовании рассматриваются различные подходы для заполнения пропусков с использованием алгоритмов машинного обучения. Для построения и тестирования моделей были взяты данные по содержанию тяжелых металлов, таких как свинец (Pb) и кадмий (Cd), так как они являются двумя наиболее часто встречающимися загрязняющими веществами [4]. В работе были использованы временные ряды концентрации свинца и кадмия по Воронежскому и Приокско-Террасному биосферным заповедникам за период с 6 декабря 1983 года по 30 сентября 2017 года.

В исследовании применяются регрессионные методы машинного обучения для заполнения непрерывных значений загрязнения почвы. В первой части используются такие алгоритмы, как k ближайших соседей (k-NN), случайный лес (RF), метод опорных векторов (SVM) и экстремальный градиентный бустинг (XGBoost). Рассматривается применение этих подходов к реальным данным о количестве тяжелых металлов с различным числом пропущенных значений – от 1 до 40%. Также считаются метрики качества RMSE (среднеквадратичная ошибка) и MAE (средняя абсолютная ошибка). В результате сравнительного анализа можно отметить, что предсказательная способность значительно снижается с увеличением недостающих значений. Тем не менее для методов XGBoost и RF она

остается на более высоком уровне, то есть ошибка растет медленнее по сравнению с k-NN и SVM. Более того, скорость обучения для всех алгоритмов относительно быстрая, поэтому наиболее точные из них вполне можно использовать для решения задач в реальном времени.

Во второй части данной работы рассматриваются архитектуры рекуррентных нейронных сетей для заполнения пропущенных значений содержания свинца и кадмия в почве. В частности, используется архитектура LSTM (Long Short-Term Memory) с различными конфигурациями гиперпараметров, таких как количество слоев, число нейронов, алгоритм оптимизации. Dropout rate выбирается в качестве случайного числа от 0 до 0,3. Нейронная сеть создается для каждой возможной архитектуры с учетом всех ограничений, в общей сложности около 150 сетей. В качестве функции потерь была выбрана среднеквадратичная ошибка. Обучение прекращается, когда прошло больше 1000 эпох или если функция потерь не уменьшалась в течение 35 эпох. Наилучшая точность, с точки зрения RMSE и MAE, достигается для конфигураций с двумя скрытыми слоями из 50 нейронов и оптимизаторами Adam / Adamax.

Предложенные в первой части исследования методы машинного обучения также могут быть успешно использованы для обработки информационных потоков в режиме реального времени [3]. Например, это может быть полезно для телекоммуникационных нагрузок или трафика. С другой стороны, нейронные сети, несмотря на достаточно медленную скорость обучения, дают более высокую точность прогнозирования. Соответственно, данный подход может найти свое применение в финансах, метеорологии, физике и медицине [2].

Литература

1. Горшенин А. К., Королев В. Ю. // Определение экстремальности объемов осадков на основе метода превышения порогового значения. Информ. и ее примен. 2018. No. 12(4). С. 16–24.
2. Akkar H., Jasim F. // Intelligent training algorithm for artificial neural network EEG classifications. Int. Journ. Intell. Syst. Appl. (IJISA). 2018. No. 10(5). P. 33–41.
3. Gorshenin A., Kuzmin V. // On an interface of the online system for a stochastic analysis of the varied information flows. AIP Conf. Proc. 1738(220009), 2016.
4. Loux N., Hassan S., Chafin C. // Empirical models of Pb and Cd partitioning using data from 13 soils, sediments, and aquifer materials. U.S. EPA, 2005.