МАТЕМАТИЧНІ МЕТОДИ ПРОГНОЗУВАННЯ ЗАБРУДНЕННЯ ПОВІТРЯ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

Кит М. О.
Науковий керівник – к.т.н., доц. Єсілевський В. С.
Харківський національний університет радіоелектроніки
61166, Харків, просп. Науки, 14, каф. прикладної математики,
тел. (057) 702-14-36, e-mail: mykyta.kyt@nure.ua

The work contains a solution of the problem of predicting air pollution by neural networks. Also, with assistance of Python programming language and Jupyter Notebook development environment, a software product was created and comparative analysis of corresponding methods and the received test result was carried out.

Проблема якості атмосферного повітря посідає особливе місце серед проблем охорони навколишнього природного середовища. Насамперед, це зумовлено великою кількістю шкідливих виробництв. Метою даної роботи є отримання прогнозу забруднення повітря з метою вжиття заходів для запобігання впливу шкідливих речовин.

Розглянемо систему обробки з кількома датчиками всередині для моніторингу в режимі реального часу Припустимо, що частота оновлення всіх датчиків ідентична і одне зі свідчень датчика є змінною, яку ми хочемо передбачити. Нехай $D=\{(x_1,t_1),\ldots,(x_n,t_n)\}$ буде набором даних показань датчика, крім цільової змінної, і нехай $y=\{(y_1,t_1),\ldots,(y_n,t_n)\}$ показники цільової змінної. Так як часовий інтервал показань ідентичний, можна опустити параметр тимчасові штампи, не викликаючи плутанини. Таким чином, набір даних D може бути перетворений як $D=\langle x_1,\ldots,x_n\rangle$, а цільова змінна $y=\langle y_1,\ldots,y_n\rangle$. Отже, існує набір даних D з набором ознак і треба отримати таку апроксимуючу функцію f, яка $f:D\to Y$, де Y мітки класів[1].

Мережі довго-короткостроковій пам'яті (Long Short Term Memory) – зазвичай просто називають "LSTM" – особливий вид RNN, здатних до навчання довгостроковим залежностям. Вони працюють неймовірно добре на великій різноманітності проблем і в даний момент широко застосовуються.

Обчислення, пов’язані з LSTM[2] :

$$i_{(t)}=\sigma\left(W_{xi}^T\cdot x_{(t)}+W_{hi}^T\cdot h_{(t-1)}+b_i\right),$$
$$c_{(t)}=f_{(t)}\otimes c_{(t-1)}+i_{(t)}\otimes g_{(t)},$$
$$y_{(t)}=h_{(t)}=o_{(t)}\otimes\tanh\left(c_{(t)}\right),$$
$$f_{(t)}=\sigma\left(W_{xf}^T\cdot x_{(t)}+W_{hf}^T\cdot h_{(t-1)}+b_f\right),$$

$$o_{(t)} = \sigma\left(W_{xo}^T \cdot x_{(t)} + W_{ho}^T \cdot h_{(t-1)} + b_o\right),$$
$$g_{(t)} = \tanh\left(W_{xg}^T \cdot x_{(t)} + W_{hg}^T \cdot h_{(t-1)} + b_g\right),$$

де $W_{xi}, W_{xf}, W_{xo}, W_{xg}$ – матриці ваг кожного з чотирьох слоїв для їх зв'язку з вхідним вектором $x_{(t)}$;

$W_{hi}, W_{hf}, W_{ho}, W_{hg}$ – матриці ваг кожного з чотирьох слоїв для їх зв'язку з попереднім короткостроковим станом $h_{(t-1)}$;

b_i, b_f, b_o, b_g – члени зміщення для кожного з чотирьох слоїв;

$h_{(t)}$ – короткостроковий стан кроку $y_{(t)}$;

f – шлюз забування;

i – вхідний шлюз;

o – вихідний шлюз.

Розглянемо види забруднень. PM10 це частинки тієї чи іншої речовини діаметром від 10 мікрометрів (мкм) і менше, PM2.5 це частинки речовини діаметром 2.5 мкм і менше. В цілому, PM2.5 можна описати як тонко дисперсні частинки. Наступним кроком буде попередня обробка даних. Для початку дані були поділені на рівні проміжки, а саме з інтервалом в одну годину. Далі усі пропущені дані були замінені на середні значення по стовпчику. Потім стовпчик з ознакою часу був перетворений на індекс, так як алгоритми нейронних мереж не можуть працювати з таким типом даних. Було побудовано декілька моделей RNN, LSTM, stacked LSTM найкраща – звичайна LSTM, тобто вона має максимальну точність, на цьому наборі даних. Також перевіряться, як модель працює на наборі test. Основними метриками в даному випадку є функція втрат та точність.

Після розрахунків результати вийшли наступні:

– simple_rnn: loss 0,02086, accuracy 85,17%;

– simple_lstm: loss 0,00519, accuracy 91,83%;

– stacked_lstm: loss 0,06461, accuracy 65,50%.

Можна сказати, що найкраща модель – це звичайна LSTM.

Для роботи на нових даних, потрібно подати набір даних за тиждень, щоб отримати прогнозовані дані на наступний тиждень, тобто модель повинна працювати постійно. Для отримання зрозумілих результатів проноза потрібно використовувати зворотне трансформування даних, тобто відмаштабувати назад.

Список використаних джерел:

1. Jinyan L., Xue L. Advaced Data Mining and Applications. Australia: ADMA, 2016. 830 p.

2. Рашка С. Python и машинное обучение. Москва: ДМК Пресс, 2017. 418 с.