

АНАЛІЗ МОЖЛИВОСТЕЙ ВИКОРИСТАННЯ МАШИННОГО ГЛИБОКОГО НАВЧАННЯ ДЛЯ РОЗПІЗНАВАННЯ СУТНОСТЕЙ

Єнгаличев С.О.

аспірант, кафедри Кібербезпеки та інформаційних технологій
Харківського національного економічного університету ім. С. Кузнеця
email: engalichev.sergiy@hneu.net
ORCID ID: 0000-0001-5298-2251

Семенов С.Г.

доктор технічних наук
Інститут безпеки та комп'ютерних наук Педагогічного університету Кракова
email: serhii.semenov@up.krakow.pl
ORCID ID: 0000-0003-4472-9234

У звіті представлено результати використання машинного глибокого навчання із застосуванням робастних М-оцінок [1] у системах аналізу та розпізнавання сутностей. У таких системах, через гетерогенність структур сутностей і систем оброблення даних [2], а також вплив зовнішніх чинників, дані, зібрані для прогнозування, можуть містити як гауссовський шум, так і випадкові викиди. У зв'язку з цим, для задач прогнозування характеристик трафіку, необхідно дослідити можливість та ефективність використання робастних оцінок максимальної правдоподібності – М-оцінок.

У рамках дослідження було запропоновано новий алгоритм, заснований на глибокому навчанні, для оцінювання продуктивності мережі (втрати пакетів) з використанням робастних М-оцінок замість традиційної оцінки середньоквадратичної помилки (MSE – Mean Square Error) у разі наявності викидів, які можуть спотворювати дані навчання. У цьому алгоритмі використовують тришарову нейронну мережу прямого поширення, де прихований шар складається з 50 прихованих нейронів.

Було створено набір даних гетерогенного трафіку, заснований на різних сутностях, з використанням стандартних інструментів сніфінгу. Цей набір даних було розділено на навчальну, тестову і перевіірочну підмножини у співвідношенні 80%, 10% і 10% відповідно. Для оцінки продуктивності було розроблено модель глибокої нейронної мережі (DNN – Deep Neural Network).

Навчання DNN було проведено з використанням робастного алгоритму зворотного поширення. У рамках дослідження було порівняно характеристики продуктивності робастної та традиційної DNN за метриками RMSE (корінь із середньоквадратичної помилки) [3] і MAPE (середня абсолютна відсоткова помилка) [4] для кожної моделі.

Для трьох випадків із різними рівнями викидів було досліджено вплив глибокого навчання та відповідної нейронної мережі DNN на продуктивність мережі. Ці випадки позначаються таким чином:

Набір А: DNN навчена на даних без шуму.

Набір В: мережа навчена на даних із включенням гаусівського шуму (GN), $G \sim N(0, 0.1)$.

Набір С: DNN навчена на даних із включенням гаусівського шуму GN, $G \sim N(0, 0.1)$, а також випадкових викидів $H1 \sim N(+15, 2)$, $H2 \sim N(-20, 3)$, $H3 \sim N(+30, 1.5)$, $H4 \sim N(-12, 3)$.

Зміна даних, що використовується в цьому випадку, визначається таким чином:

Дані = $(1-\varepsilon) G + \varepsilon(H1+H2+H3+H4)$, де ε становить 10% від даних.

Для цього експерименту було згенеровано набір даних на основі моделі мережі банківської установи. Потім, точки даних у координатах x і y були піддані спотворенню за допомогою гауссовського шуму з нульовим середнім значенням і стандартним відхиленням 0.1, $G \sim N(0, 0.1)$. Змінна ε точок даних була випадково обрана, а потім замінена з імовірністю ε фоновим шумом, рівномірно розподіленим у заданому діапазоні.

DNN-архітектура являє собою багатошарову нейронну мережу прямого поширення MFNN, що складається з трьох шарів, включно з одним прихованим шаром зі 100 прихованими нейронами. Навчання мережі здійснювалося з використанням робастного алгоритму зворотного поширення BP (Back Propagation) [5] і застосуванням функцій втрат, згаданих раніше. У моделі глибокого навчання DL (Deep Learning) було обрано найбільш підходящі архітектурні параметри, як-от розмір пакета, кількість епох, функція активації, функція втрат, швидкість навчання і мінімальні втрати.

У цьому дослідженні використовували розмір пакета 32, 5000 епох, функцію навчання Traincgrf, швидкість навчання 10^{-2} і мінімальні втрати 10^{-2} . Для всіх нейронів у прихованих шарах було обрано сигмоїдну функцію активації (Tansig), а для нейрона у вихідному шарі було обрано лінійну функцію активації.

Для забезпечення оптимальної продуктивності DNN необхідно нормалізувати вхідні дані, щоб вони перебували в інтервалі $[-1, 1]$, що відповідає фактичним мінімальним і максимальним значенням. DNN використовує пропускну спроможність як вхідні дані і втрати пакетів як вихідні дані (прогнозований результат) для оцінки оптимальної продуктивності мережі на основі зібраного набору даних. Основна мета полягає в розробці робастної нейронної мережі DNN, яка здатна оцінювати продуктивність мережі за умов впливу зовнішніх факторів і різноманітних умов.

Для досягнення цієї мети було проведено порівняльне дослідження робастних і традиційних DNN з використанням метрик середньоквадратичної помилки RMSE і середньої абсолютної помилки у відсотках MAPE. Метою дослідження було визначити, які оцінки демонструють найкращі результати для аналізованого додатка.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (t_i - y_i)^2}{N}}$$

$$MAPE = \frac{50}{n} \sum_{i=1}^N \left| \frac{(t_i - y_i)}{t_i} \right|$$

Дослідження показали, що використання робастної оцінки та робастної оцінки Коші призводить до збільшення середньоквадратичної помилки більш ніж у п'ять разів, тоді як абсолютна помилка є приблизно в 1,5 рази меншою у порівнянні з використанням методу найменших квадратів.

Список літератури:

1. D.Q.F. de Menezes, D.M. Prata, A.R. Secchi, J.C. Pinto, A review on robust M-estimators for regression analysis, Computers & Chemical Engineering, Volume 147, 2021, 107254, ISSN 0098-1354, <https://doi.org/10.1016/j.compchemeng.2021.107254>.
2. S. Semenov, V. Lymarenko, S. Yenhalychev and S. Gavrilenko, "The Data Dissemination Planning Tasks Process Model Into Account the Entities Differentity," 2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT), Athens, Greece, 2022, pp. 1-6, doi: 10.1109/DESSERT58054.2022.10018695.
3. Ossai, Chinedu. (2019). A Data-Driven Machine Learning Approach for Corrosion Risk Assessment – A Comparative Study. Big Data and Cognitive Computing. 3. 28. 10.3390/bdcc3020028.
4. Anagnostis, Athanasios & Papageorgiou, Elpiniki & Bochtis, Dionysis. (2020). Application of Artificial Neural Networks for Natural Gas Consumption Forecasting. Sustainability. 12. 6409. 10.3390/su12166409.
5. Li, J., Cheng, Jh., Shi, Jy., Huang, F. (2012). Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement. In: Jin, D., Lin, S. (eds) Advances in Computer Science and Information Engineering. Advances in Intelligent and Soft Computing, vol 169. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30223-7_87