ОСНОВНІ МОДЕЛІ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ НАВАНТАЖЕННЯ НА СЕРВЕРИ ТА ПОТРЕБ У МАСШТАБУВАННІ

Чиж О.А.
Національний технічний університет "Харківський політехнічний інститут", м. Харків

Прогнозування навантаження на сервери є ключовою складовою адаптивного управління ресурсами. Для цього використовуються різні моделі машинного навчання: регресійні методи, рекурентні нейронні мережі (RNN), а також трансформери. Ці моделі аналізують історичні дані, такі як обсяги трафіку, використання процесорних потужностей, пам'яті та інших серверних ресурсів, і генерують прогнози щодо майбутніх потреб у масштабуванні. Для підвищення точності прогнозів, окрім традиційних статистичних методів, активно застосовуються методи глибинного навчання, які здатні ефективно працювати з великими обсягами даних та складними залежностями між різними параметрами навантаження.

Ефективність моделей залежить від якості даних, алгоритмів навчання, обчислювальних потужностей, а також здатності моделі адаптуватися до нових умов. Наприклад, використання моделей Long Short-Term Memory (LSTM) показує високу точність для задач із часовими рядами, оскільки ці моделі можуть враховувати довгострокові залежності в даних. Водночас, трансформери, завдяки своїм можливостям обробки великих обсягів даних і паралельній обробці, мають перевагу в обробці даних із складними залежностями, що дозволяє їм досягати високих результатів на великих наборах даних [1]. Сучасні гібридні підходи комбінують ці моделі з іншими техніками, такими як ансамблювання та мета-навчання, для досягнення ще вищої точності прогнозів [2].

Оцінка таких моделей дозволяє обрати оптимальне рішення для конкретних сценаріїв масштабування.

Література:

1. Neural Networks and Deep Learning / Charu C. Aggarwal – Springer Cham, 2023. – 529 p.
2. Ensemble Methods Foundations and Algorithms / Zhi-Hua Zhou – Chapman & Hall, 2012. – 236 p.