

УДК 519.65

А.О. Дорошенко, студент гр ПГ-01мп
КПІ ім. Ігоря Сікорського

АНАЛІЗ ПРОБЛЕМ СУЧАСНИХ МОДЕЛЕЙ У АВТОМАТИЧНИХ СИСТЕМАХ ПРОГНОЗУВАННЯ

Анотація. Аналізуються причини необхідності використання ускладнених методів прогнозування часових рядів. Досліджується проблема прогнозування виробництва приладів на підставі даних споживання товарів. Розглядаються існуючі методи машинного навчання. Визначаються основні недоліки наведених. Сформовано основні вимоги для створення автоматизованої системи прогнозування попиту, яка дозволить ефективніше керувати виробництвом у реальному часі.

Ключові слова: прогнозування, часові ряди, машинне навчання, модель, автоматична система.

ВСТУП

З появою нових способів створення та поширення інформації зростання потоку документів збільшувалося в рази. За прогнозами International Data Corporation (IDC) кількість даних на планеті подвоюватиметься кожні 2 роки, причому частка корисної інформації складатиме всього 35% від усієї згенерованої, а обсяг загальносвітових даних зросте з 33 зеттабайт у 2018 році до 175 у 2025 [1]. З урахуванням стрімкого зростання кількості інформації, правильне зберігання та управління великою кількістю даних все більше необхідно для аналізу та управління великими організаціями чи виробництвами. Розуміння споживання кількості виробленого товару дозволяє компаніям ефективно і своєчасно підлаштовуватися під поточний попит на товар, знижуючи або збільшуючи виробничі потужності.

Проте, сучасні автоматичні системи та моделі прогнозування не дозволяють належним чином реагувати на непередбачувані події чи події «непереборної сили». Прикладом таких подій може бути обставини пов'язані з Covid-19 і кризою мікрочіпів, за яких виробники не мали можливості спрогнозувати і не були готові відреагувати на підвищений попит [2]. Також вплив таких подій негативно впливає на збір і аналіз статистичних даних та наступне їхнє прогнозування.

ОГЛЯД НАЯВНИХ АКТУАЛЬНИХ РІШЕНЬ

За визначенням, зміну попит товару можна охарактеризувати як нестационарний часовий ряд. При дослідженні стаціонарних випадкових процесів кількість значень, що спостерігаються, безпосередньо впливає на якість прогнозування або визначення характеристик розподілу випадкової величини. Однак це правило не виконується для нестационарних процесів, оскільки нестационарність не дозволяє отримати достатню кількість значень величин часового ряду для побудови математичної моделі. Інакше кажучи, характеристики випадкової величини нестационарного процесу неспроможні давати точний прогноз та подальшу поведінку через виникнення моментів розладки, які можуть бути викликані непередбачуваними подіями [3].

Варто зазначити, що економічні процеси – одна з найбільш вивчених галузей аналізу та прогнозування часових рядів. Цьому сприяє велика зацікавленість учасників отримання позитивних результатів. Наприклад, у статті

[4] був описаний підхід до фундаментального аналізу даних часових рядів для виявлення зв'язків між поведінкою ринку та зовнішнім середовищем. Авторами статті було проведено дослідження за допомогою нейронної мережі PSI (PSI Sigma) [5] щодо існування взаємозв'язків між зовнішнім середовищем та валютною парою USD/GBR. В результаті проведення робіт авторами було показано здатність даної НМ до більш точної фіксації кореляції. Поєднання з алгоритмом Gene Expression допомогло досягти більш точного прогнозування порівняно з класичними методами та моделями. В іншій роботі [6], авторами стверджується, що нейронні мережі можуть добре справлятися з моделюванням руху цін фінансових індексів. Використання параметричних НМ з активними нейронами спільно з непараметричними моделями, що самоорганізуються, дозволяють отримувати більшу кількість інформації з вхідних даних і може підходити для щоденного прогнозування.

Одним із головних недоліків багатьох підходів є застосування єдиної реалізації моделі випадкового процесу. У разі використання єдиної моделі для прогнозування деякої кількості схожих за своїми характеристиками часових рядів, необхідно використання методів, що базуються на еволюції сукупності розподілів процесу.

Одним із найпопулярніших сучасних методів є метод XGBoost. З 29 рішень-переможців, опублікованих у блозі Kaggle у 2015 році, 17 рішень використали XGBoost. Успіх системи був також засвідчений у KDDCup 2015, де XGBoost використовувався кожною командою з топ-10 [7]. Алгоритм спочатку стартував як дослідницький проект Тяньцзі Чена (Tianqi Chen) і замислювався як додаток, який можна було налаштувати за допомогою конфігураційного файлу. В основі XGBoost лежить алгоритм градієнтного бустингу дерев рішень – техніка машинного навчання задач класифікації та регресії [7]. Особливістю такої моделі є побудова моделі прогнозування у формі ансамблю слабких передбачуваних моделей, що часто використовуються дерев рішень. Ітеративне навчання дозволяє навчитися новій моделі на підставі даних попередньої, а також заздалегідь розрахованих відхилень прогнозу від факту. Найбільш важливим фактором успіху XGBoost є його масштабованість у всіх сценаріях. Система працює більш ніж у десять разів швидше за існуючі популярні рішення на одній машині, і масштабується до мільярдів прикладів у розподілених налаштуваннях або налаштуваннях з обмеженням пам'яті [7]. Проте суттєвим недоліком такої моделі є можливість до перенавчання – модель пристосовується до прогнозування строго певного паттерну чи поведінки часового ряду. Альтернативним рішенням може виступати сучасні архітектури рекурентних нейронних мереж, моделі яких меншою мірою схильні до впливу перенавчання, здатні до повторного навчання на підставі нових вхідних даних і вже наявних ваг попередньої моделі, а також до ітеративного прогнозування або прогнозування послідовності.

ОСНОВНІ ПРОБЛЕМИ СУЧАСНИХ МОДЕЛЕЙ

При проектуванні автоматичних моделей прогнозування часових рядів, найчастіше, виникає серія проблем, з якими інженер намагається вирішити

різними способами. Ці проблеми можна поділити на декілька основних частин: інфраструктура, програмне забезпечення, проблеми аналізу часових рядів, проблеми повторного навчання моделей новим даними та проблеми перенавчання.

При виборі моделі, насамперед, потрібно визначитися з доцільністю використання обчислювальної техніки з необхідним обсягом інформації. Найбільший час обчислень, найчастіше, займає навчання моделі і може займати від десятка хвилин до декількох місяців. Сучасні архітектури GPU здатні до паралельного обчислення, наприклад, однією з найпоширеніших технологій є Nvidia Cuda, що може значно прискорити навчання.

Правильно вибране програмне забезпечення може значної мірою підвищити швидкість роботи автоматизованих систем. За останні роки популярні мови будування моделей, таких як R або Python, починають поступатись іншим, які можуть надати більшу гнучкість у використанні. Прикладом такої мови може виступати сучасний Go, який дозволяє використовувати вже звичні фахівцям бібліотеки з більшою швидкістю і ще більшою простотою.

Більшість фахівців поділяють основні проблеми даних при прогнозуванні часових рядів на дві частини: пропуски даних та викиди. І хоча існує велика кількість методів боротьби з пропусками та викидами, при використанні автоматизованих систем, краще уникати цих проблем. Основною вимогою до будування автоматичних систем є можливість завантаження цілісних та максимально повних даних з певним постійним кроком.

Як було раніше зазначено, для побудови автоматичної системи необхідним є вибір моделі, яке зможе забезпечити можливість повторного навчання при отриманні нових даних. При цьому, обрана модель не повинна недоліку перенавчання тому, що накопичення похибки буде збільшуватись з кожною ітерацією навчання та прогнозу. Найпопулярнішими рішеннями з даними вимогами можуть бути моделі машинного навчання XGBoost та моделі глибинного навчання – специфічні архітектури НМ.

ВИСНОВОК

З аналізу джерел можна зробити висновок, що будування автоматичних систем прогнозування залежить від обсягу наявних даних та методу їх отримання. Основною вимогою є цілісність та відсутність пропусків даних, а також можливість інкрементального завантаження. Використання моделей залежить від наявних даних, але слід звернути увагу на моделі XGBoost або деяких типів нейронних мереж. Використання таких моделей зумовлено їх можливостям до постійного навчання новими даними. Інфраструктурне та програмне рішення може залишатись на вибір фахівцю.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- [1] David Reinsel, John Gantz, John Rydning, «Data Age 2025: The Digitization of the World», [Електронний ресурс]. – Режим доступу: <https://www.seagate.com/ru/ru/our-story/data-age-2025>

- [2] Businessinsider, Global chip shortages are expected to cost automakers \$210 billion in 2021 — almost double previous estimates, a consulting firm says [Електронний ресурс]. – Режим доступу: <https://www.businessinsider.com/global-chip-shortage-supply-chain-cost-auto-industry-210-billion-2021-9>
- [3] Шмойлова Р. А. Общая теория статистики: Учебник. — М.: Финансы и статистика, 2002.
- [4] Khadjeh Nassirtoussi A., Ying Wah T., Ngo Chek Ling D. A novel FOREX prediction methodology based on fundamental data. Afr. J. Bus. Manag. 2011;5(20):8322-8330, DOI:10.5897/AJBM11.798
- [5] Swapna Rekha, Janmenjoy Nayak, H. S. Behera Pi-Sigma Neural Network: Survey of a Decade Progress H. [Електронний ресурс]. – Режим доступу: https://link.springer.com/chapter/10.1007/978-981-15-2449-3_37
- [6] Anastasakis L., Mort N. Exchange rate forecasting using a combined parametric and non-parametric selforganising modelling approach. Expert Sys. Appl. 2009;36(10):12001-12011
- [7] Tianqi Chen , Carlos Guestrin XGBoost: A Scalable Tree Boosting System, 2016. [Електронний ресурс]. – Режим доступу: <https://arxiv.org/pdf/1603.02754.pdf>
- Наук. керівник – зав. каф., д.т.н., проф. Бурау Н.І***