Шаховська Христина Романівна

Національний університет “Львівська політехніка”, Львів

ORCID 0000-0002-9914-229X

**Повторюваний концептуальний дрейф: виклики та адаптація**

У сучасних реаліях машинне навчання стикається з низкою викликів, які пов’язані із змінністю даних. Більшість традиційних моделей працюють, спираючись на припущення, що розподіл даних є статичним. Однак у реальних сценаріях це рідко відповідає дійсності. Потокові дані, наприклад, часто змінюють свої закономірності з часом, що породжує явище, відоме як концептуальний дрейф [1]. Концептуальний дрейф проявляється в різних формах: раптові зміни в даних, поступові зрушення або навіть періодичне повторення раніше спостережуваних патернів.

Ще складнішою є ситуація з повторюваним концептуальним дрейфом [2], коли вже знайомі моделі даних знову з’являються через певний час. Це особливо характерно для сезонних тенденцій, змін уподобань користувачів або періодичних процесів у системах. Для таких сценаріїв потрібні моделі, які не тільки здатні виявляти зміни, але й швидко до них адаптуватися, забезпечуючи стабільність і високу точність роботи. Наприклад, сезонні зміни у фінансових даних, циклічні коливання в уподобаннях споживачів чи періодична поведінка технічних систем ставлять моделі перед необхідністю не лише адаптуватися до змін, але й ефективно використовувати раніше накопичені знання.

Нездатність моделі ефективно працювати з концептуальним дрейфом може призводити до серйозних наслідків: зниження точності прогнозів, неправильні класифікації чи навіть значні фінансові втрати. Наприклад, у системах виявлення шахрайства або медичного моніторингу пропуск змін у даних може мати критичні наслідки.

Щоб ефективно працювати з концептуальним дрейфом, сучасні моделі повинні володіти кількома ключовими якостями.[3] По-перше, вони мають бути здатні швидко розпізнавати зміни у даних, зокрема повторювані. По-друге, важливо забезпечити баланс між стабільністю (збереженням попередніх знань) і пластичністю (здатністю адаптуватися до нових патернів). Це вимагає ефективного управління пам’яттю та механізмів для виділення релевантних даних із минулого.

Зазвичай для вирішення проблем концептуального дрейфу використовують онлайн машинне навчання. Онлайн машинне навчання дозволяє моделювати потоки даних у реальному часі, враховуючи нестабільності та зміни у даних (дрейфи, зсуви). Однак основні виклики, які виникають[4]:

* Стабільність проти адаптивності: Моделі часто втрачають попередні знання через необхідність адаптації до нових змін, що призводить до потреби в повторному навчанні.
* Необхідність адаптації до концептуальних дрейфів: Зміни в статистичних властивостях даних, як поступові, так і різкі, ускладнюють підтримку точності моделі.
* Неефективність ресурсів: Постійне навчання з нуля потребує великих витрат обчислювальних ресурсів і часу, особливо у сценаріях, де старі патерни стають знову актуальними.

Враховуючи вищезгадані проблеми зараз активно досліджуються адаптивні алгоритми. Головна відмінність між інкрементальними та адаптивними алгоритмами полягає у стратегіях забування. Адаптивні алгоритми використовують явні механізми для ігнорування нерелевантної інформації, тоді як інкрементальні алгоритми просто додають нові знання до існуючих. Фактично, адаптивні алгоритми можна вважати «розширеними інкрементальними алгоритмами»[5], які можуть адаптуватися до змін у потоці даних.

Однак динамічний характер потокових даних викликає низку проблем на етапах зберігання та навчання. Генеративні процеси у часових рядах можуть змінюватися з часом, а моделі, натреновані на старих даних, зазвичай втрачають точність при таких змінах. Вирішенням цих проблем є розробка механізмів, які можуть ефективно працювати з концептуальними дрейфами, зберігаючи продуктивність у періоди стабільності[6].

Розглянемо кілька підходів до вирішення цієї проблеми. У контексті концептуальних дрейфів одним із найбільш продуктивних рішень стали ансамблі, які дозволяють комбінувати кілька моделей для підвищення точності та масштабованості. Вони добре інтегруються у розподілені обчислювальні системи та масштабуються для роботи з великими потоками даних. Однак покроковий характер онлайн-навчання створює обчислювальні труднощі для паралельних моделей. У таких випадках адаптивні підходи, як-от стратегія навчання на міні-батчах[7], допомагають оптимізувати процес, дозволяючи використовувати алгоритми офлайн-рівня для динамічних середовищ і нестабільних даних.

Інший підхід — прееквентне оцінювання(prequential or interleaved test-then-train)[8]: Ефективний метод моніторингу продуктивності, де кожен екземпляр даних використовується для прогнозу, оцінки помилки та подальшого навчання моделі. У цьому підході, тести виконуються на нових даних перед використанням для навчання моделі. Це забезпечує ефективніше використання даних порівняно з традиційним підходом "розділення на навчальні та тестові набори".

Мета-навчальні підходи для поточних потоків даних дозволяють повторно використовувати попередньо навчені моделі лише після виявлення зміни в даних. Це забезпечує ефективність через використання міні-батчів і збереження старих моделей для відповідних дрейфів. У зв'язку з цим, були запропоновані різні методи, зокрема ансамблі з активним виявленням дрейфу та використанням історії попередніх моделей[9]. Для цього набір отриманих даних зберігається в буфері для кожного класифікатора та порівнюється з вхідним потоком у разі дрейфу концепції.

Новітні методи включають використання нейронних мереж для передбачення типу дрейфу, що дозволяє адаптувати онлайн-алгоритми до нових змін у даних [10]. Автори використовують попередньо навчений класифікатор, щоб навчитися виявляти різні типи дрейфу в потоці даних. Автори пропонують нейронну мережу, яка вивчає прототипи(набір екземплярів, створених або вибраних алгоритмом) для кожного класу дрейфу, включаючи базовий клас «без дрейфу». Після цього детектор дрейфу можна використовувати як частину будь-якого онлайн-навчального алгоритму, маючи перевагу в тому, що він може передбачити виникнення дрейфу та його тип.

Онлайн машинне навчання має великий потенціал для роботи з потоковими даними, але його ефективність залежить від здатності моделей адаптуватися до змін, зберігати попередні знання та оптимізувати використання ресурсів. Адаптивні алгоритми є ключем до вирішення цих викликів, пропонуючи сучасні підходи до роботи з концептуальними дрейфами та повторюваними патернами. Найскладніше при роботі з повторюваним концептуальним дрейфом — це розрізнити, коли дані сигналізують про повернення до старих патернів, а коли з’являється щось принципово нове. Помилки в таких випадках можуть призводити до некоректних рішень. Саме тому важливим напрямком досліджень є розробка ефективних алгоритмів, здатних розпізнавати такі ситуації. У перспективі вирішення цих викликів дозволить створити більш стійкі та адаптивні системи, які можуть працювати з великими потоками даних у режимі реального часу. Це відкриє нові можливості для різних галузей і допоможе підвищити ефективність прийняття рішень навіть у найбільш складних умовах.

**Література**

1. TSYMBAL, Alexey. The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin, 2004, 106.

2. SUÁREZ-CETRULO, Andrés L.; CERVANTES, Alejandro; QUINTANA, David. Incremental market behavior classification in presence of recurring concepts. Entropy, 2019, 21.1: 25.

3. WARES, Scott; ISAACS, John; ELYAN, Eyad. Data stream mining: methods and challenges for handling concept drift. SN Applied Sciences, 2019, 1: 1-19.

4. ELWELL, Ryan; POLIKAR, Robi. Incremental learning of concept drift in nonstationary environments. IEEE transactions on neural networks, 2011, 22.10: 1517-1531.

5. GAMA, João, et al. A survey on concept drift adaptation. ACM computing surveys (CSUR), 2014, 46.4: 1-37.

6. LU, Jie, et al. Learning under concept drift: A review. IEEE transactions on knowledge and data engineering, 2018, 31.12: 2346-2363.

7. CASSALES, Guilherme, et al. Improving the performance of bagging ensembles for data streams through mini-batching. Information Sciences, 2021, 580: 260-282.

8. CERQUEIRA, Vitor; TORGO, Luis; MOZETIČ, Igor. Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning, 2020, 109.11: 1997-2028.

9. GONÇALVES JR, Paulo Mauricio; DE BARROS, Roberto Souto Maior. RCD: A recurring concept drift framework. Pattern Recognition Letters, 2013, 34.9: 1018-1025.

10. U, Hang, et al. Meta-ADD: A meta-learning based pre-trained model for concept drift active detection. Information Sciences, 2022, 608: 996-1009.