УДК 004.94 https://doi.org/10.15407/pp2025.02.112

Я. Жилюк, В. Плескач

ГІБРИДНИЙ ПІДХІД ДО ОБРОБЛЕННЯ НЕПОВНИХ ПОТОКОВИХ ДАНИХ У РОЗПОДІЛЕНИХ СИСТЕМАХ РЕАЛЬНОГО ЧАСУ

У статті розглянуто проблему оброблення неповних потокових даних у розподілених інформаційних системах реального часу, зокрема у контексті інтелектуального аналізу даних. Зазначено, що традиційні методи імпутації є малоефективними в умовах обмежених ресурсів, високих вимог до швидкості оброблення та динамічного характеру потоків. Запропоновано гібридний підхід, що поєднує *федеративне* навчання, контекстну імпутацію та адаптацію до концептуального дрейфу. Метод дозволяє локальним розподіленим обчислювальним вузлам тренувати полегшені моделі імпутації на власних даних із подальшою централізованою агрегацією, зворотним поширенням глобальної моделі та її динамічним оновленням. Експериментальна перевірка на реальному датасеті показала переваги підходу за точністю (RMSE, MAE) та навантаженням на мережу порівняно з базовими методами. Отримані результати засвідчують ефективність запропонованого методу в умовах розподілених середовищ з обмеженими обчислювальними ресурсами.

Y. Zhyliuk, V. Pleskach

HYBRID APPROACH TO PROCESSING INCOMPLETE STREAM DATA IN DISTRIBUTED REAL-TIME SYSTEMS

The article considers the problem of processing incomplete streaming data in distributed real-time systems, in particular in the context of data mining. It is noted that traditional methods of imputation are ineffective in conditions of limited resources, high requirements for processing speed and dynamic nature of streams. A hybrid approach combining federated learning, contextual imputation and adaptation to conceptual drift is proposed. The method allows local distributed computing nodes to train lightweight imputation models on their own data, followed by centralised aggregation, backpropagation of the global model and its dynamic updating. Experimental verification on a real dataset has shown the advantages of the approach in terms of accuracy (RMSE, MAE) and network load compared to the baseline methods. The obtained results prove the effectiveness of the proposed method in distributed environments with limited computing resources.

Вступ

З поширенням розподілених систем реального часу (РСРЧ) зростає рівень впровадження інтелектуального аналізу потокових даних у багатьох доменах, зокрема, Інтернеті речей (IoT), сенсорних мережах та фінансових ринках.

Потоки даних вирізняються своєю безперервною та необмеженою природою, чутливістю до часу та часто великим обсягом, що створює проблеми під час керування та оброблення даних [1].

Розподіл джерел даних ще більше ускладнює ці виклики, створюючи складнощі з боку синхронізації даних, інтеграції до загальної інфраструктури розподілених інформаційних систем. Значною перешкодою для ефективного використання інформації, що міститься в цих розподілених потоках даних, є поширене явище неповних або відсутніх даних. Ця проблема виникає внаслідок різних факторів, таких як несправності обладнання, помилки мережі та непередбачуваність процесів збору даних.

Причини неповноти даних можуть варіюватися від недостатньої повноти збору даних до неузгодженості в методологіях, неточностей у вимірюваннях, проблем із достовірністю чи цілісністю даних або навіть розбіжностей у часових межах збору даних. Крім того, відсутність даних може бути зумовлена різними механізмами, зокрема, повністю випадковою відсутністю

(*M*issing *C*ompletely *A*t *R*andom, MCAR), випадковою відсутністю (*M*issing *A*t *R*andom, MAR), невипадковою відсутністю (*M*issing *N*ot *A*t *R*andom, MNAR) та структурною відсутністю даних, кожен з яких має свої наслідки для функціонування системи й процесу аналізу потокових даних [2].

Традиційні підходи до імпутації даних орієнтовані здебільшого на статичні вибірки, виявляються малоефективними в умовах розподілених потокових систем реального часу. Методи імпутації – це прийоми, які використовують для заповнення відсутніх або пропущених значень у даних на основі контексту, для забезпечення їхньої повноти для подальшого аналізу чи побудови моделей.

Їхня неспроможність адаптуватися до динамічної природи потокових даних зумовлює проблеми масштабування, а також затримки під час обробки в розподілених обчислювальних середовищах. У зв'язку з цим виникає потреба у розробці нових методів імпутації, здатних враховувати специфіку розподілених потоків даних і забезпечувати ефективність інтелектуального аналізу в умовах жорстких часових обмежень.

Метою цієї статті є розроблення та обґрунтування гібридного підходу до імпутації неповних потокових даних у розподілених інформаційних системах реального часу, що враховує динамічний характер даних, обмежені обчислювальні ресурси та потребу в адаптації до концептуального дрейфу для підвищення ефективності інтелектуального аналізу даних. Зростаюча залежність від розподілених систем для збору й аналізу даних підкреслює критичність вирішення проблем із відсутніми даними в цих потоках, щоб розкрити їхній повний потенціал для отримання значущої інформації в процесі інтелектуального аналізу.

Проблема неповних даних у розподілених потоках для інтелектуального аналізу

Наявність неповних даних у розподілених потоках значно перешкоджає ефективності задач інтелектуального аналізу. Ці проблеми проявляються по-різному в розпізнаванні образів, виявленні аномалій і прогнозному моделюванні, кожне з яких вимагає ретельного розгляду та індивідуальних рішень.

Алгоритми, призначені для ідентифікації кластерів або класифікації даних, значною мірою покладаються на повний набір атрибутів, для точного визначення подібності та відмінності аналізованих даних. У разі відсутності певної частини інформації, вказані алгоритми можуть менш точно ефективно виконувати задачі класифікації, що призводить до зниження здатності розпізнавати значущі патерни. Наприклад, у мережах датчиків, розгорнутих у географічній зоні, відсутність показань від певних датчиків може призвести до неповних просторових або часових моделей, що ускладнює розуміння відстежуваного явища. Проблема відсутніх даних може посилюватися в поєднанні з незбалансованими наборами даних, де недостатнє представлення певних класів об'єктів може додатково посилюватися неповнотою інформації, що призводить до неправильної інтерпретації даних моделями машинного навчання [3].

Відсутність даних може призвести до виявлення помилкових закономірностей або упущення справжніх закономірностей, що зрештою призведе до некоректного розуміння основних явищ.

Виявлення аномалій у даних під час інтелектуального аналізу потоків у РСРЧ значно ускладнюється неповнотою вхідних даних. Аномальні точки даних, які представляють відхилення від норми, можуть бути замасковані чи неправильно інтерпретовані як відсутні значення, або, навпаки, самі відсутні значення можуть бути позначені як аномалії.

Ігнорування відсутніх значень є поширеною стратегією попереднього оброблення даних, що становить ризик згладжування або спотворення справжніх аномалій, тим самим нівелює саму мету виявлення відхилень. У розподілених середовищах, де дані можуть надходити асинхронно та з різним ступенем повноти з різних вузлів моніторингу, завдання виявлення аномалій стає ще більш складним. Наприклад, під час моніторингу мережевого трафіку відсутня

інформація про пакети може або приховати шкідливу активність, яка виглядатиме як аномалія, якщо вона завершена, або хибно передбачити вторгнення через неповну передачу [4].

Прогнозування в контексті неповних потоків даних також створює значні перешкоди. Відсутні значення в навчальних даних, які використовуються для побудови прогнозних моделей, можуть призвести до упереджених або неефективних моделей, що зрештою вплине на їхню точність і надійність. Коли ці моделі застосовують до потоків в РСРЧ, які також містять відсутні дані, точність отриманих прогнозів може бути суттєво знижена. Проблема ще більше ускладнюється явищем зсуву інформації, яке є поширеним у потокових даних, де основний розподіл даних змінюється з часом [11].

Прогнозні моделі, базовані на навчанні на історичних даних із пропущеними значеннями, можуть виявитися недостатньо адаптованими до змін у поточному потоці даних, що може призвести до поступового зниження точності прогнозів під час розвитку процесу. Наприклад, у закладах охорони здоров'я система моніторингу показників життєдіяльності пацієнтів у реальному часі, яка спирається на прогнозні моделі, може генерувати неточні прогнози критичних подій, якщо у вхідному потоці даних відсутні важливі показники. Прогнозні моделі, навчені на неповних даних, можуть вивчити помилкові зв'язки між даними або не зафіксувати базову динаміку, що призведе до ненадійного узагальнення та неточних прогнозів на нових, потенційно також неповних даних. Відсутні значення в навчальних даних зменшують обсяг інформації, доступної для навчання, потенційно призводячи до надмірного, недостатнього підбору чи зміщення параметрів моделі.

Обмеження традиційних методів доповнення даних для розподілених потоків у реальному часі

Традиційні методи доповнення даних, хоча й ефективні в певних контекстах, часто виявляються не ефективними для оброблення відсутніх даних у потоках розподілених даних у режимі реального часу, особливо в умовах застосування методів інтелектуального аналізу, що пояснюється кількома факторами, зокрема, затримкою, обчислювальними обмеженнями та мінливим характером розподілу даних.

Одним з основних обмежень є затримка. Багато традиційних методів доповнення, такі як множинна імпутація та складні підходи на основі машинного навчання, є обчислювально ресурсомісткими та можуть призводити до значних затримок в обробленні. Оброблення потоків у реальному часі за своєю природою вимагає доповнення даних з низькою затримкою, щоб забезпечити своєчасний аналіз та ухвалення рішень. Навіть регресійна імпутація, яка може бути відносно точною, може призводити до неприйнятної затримки, особливо якщо базові регресійні моделі є складними або потребують частого оновлення. Потреба в негайному обробленні потоків даних суперечить притаманній затримці багатьох точних традиційних методів імпутації, змушуючи йти на компроміс між точністю та швидкістю. Застосунки реального часу вимагають швидкого аналізу. Методи імпутації, які потребують значного часу оброблення на централізованому вузлі, стають вузькими місцями в розподіленому потоковому середовищі, затримуючи загальний аналіз і потенційно зменшуючи релевантність результатів аналізу.

Обчислювальні обмеження також створюють значну проблему. Розподілене оброблення потоків часто відбувається на пристроях з обмеженими ресурсами, таких як датчики Інтернету речей та периферійні сервери. Багато складних методів імпутації вимагають значної обчислювальної потужності та ресурсів пам'яті, які можуть перевищувати можливості цих пристроїв. Для ефективного завершення оброблення даних у таких середовищах необхідні легкі та ефективні методи імпутації. Розгортання обчислювально ресурсомістких методів імпутації на периферії розподіленої системи обробки потоків часто є неможливим через обмежені ресурси, доступні на цих пристроях. Потоки даних часто надходять із численних

пристроїв з низьким енергоспоживанням. Алгоритми імпутації мають бути достатньо ефективними, щоб функціонувати локально або на периферії, не споживаючи надмірної енергії або обчислювальної потужності, що може вплинути на основну функцію та строк служби пристрою [6].

Традиційні методи доповнення часто припускають, що основні статистичні властивості даних залишаються постійними з часом, однак потоки даних є за своєю суттю динамічними, а їхні розподіли змінюються, коли виникають нові закономірності, а старі зникають. Моделі імпутації, навчені на знімку історичних даних, можуть з часом ставати неточними у процесі розвитку потоку даних, що призводить до низької продуктивності у врахуванні відсутніх значень. Для вирішення цієї проблеми потрібні адаптивні методи імпутації, які можуть виявляти та коригуватися до концептуального дрейфу даних [11]. Дрейф даних (data drift) – це зміна статистичних властивостей даних, які використовують у моделях машинного навчання, протягом певного часу. Це явище може призвести до зниження ефективності моделі, оскільки алгоритм, навчений на певних даних, може стати менш точним або некоректним, якщо дані, на яких вона працює, змінюють. Динамічна природа потоків даних з їхніми концептуальними закономірностями та розподілами, що розвиваються, вимагає методів імпутації, спроможних адаптуватися в режимі реального часу для підтримки точності [7].

Зрештою традиційні методи доповнення несуть потенціал для внесення систематичних помилок у дані. Прості методи, такі як імпутація середнього значення, хоча й легко реалізувати, можуть спотворювати розподіл основних даних і недооцінювати дисперсію, що потенційно може призвести до хибно позитивних висновків у подальшому аналізі. Регресійна імпутація, якщо обрана модель регресії неточно відображає справжні зв'язки між змінними, також може призвести до систематичної помилки. Ефективність будь-якого методу імпутації тісно пов'язана з механізмом, через який дані відсутні (MCAR, MAR, MNAR), а ігнорування цього механізму може призвести до систематичних помилок імпутації. Систематична помилка, внесена під час процесу доповнення, може потім поширюватися та негативно впливати на результати подальших завдань інтелектуального аналізу. Неправильно застосовані методи доповнення можуть вносити систематичні помилки в дані, що потенційно може призвести до помилкових висновків та рішень на основі подальшого інтелектуального аналізу. Мета імпутації – заповнити відсутні значення правдоподібними оцінками. Однак, якщо метод імпутації робить неправильні припущення щодо даних або механізму відсутності, це може спотворити справжні основні зв'язки між даними, які поширюються по всьому ланцюгу інтелектуального аналізу [7].

Визнаючи обмеження традиційних методів імпутації в контексті потоків даних, дослідники розробили різні методології, спеціально спрямовані на вирішення проблем, що виникають через неповні дані в цих динамічних середовищах. Ці підходи варіюються від простих методів видалення до більш складних методів імпутації, адаптованих до унікальних характеристик потокових даних.

Методи видалення представляють собою простий підхід до обробки відсутніх даних. Спискове видалення передбачає видалення всіх спостережень з набору даних, якщо вони містять одне чи декілька відсутніх значень. Хоча цей метод простий у реалізації, він може призвести до значної втрати цінної інформації, особливо коли рівень відсутніх даних високий. Більше того, якщо відсутні дані не є повністю випадковими (MCAR), спискове видалення може внести зміщення в подальший аналіз. З іншого боку, попарне видалення намагається зменшити втрату інформації, використовуючи всі доступні дані для кожного конкретного аналізу. Це означає, що аналіз може базуватися на різних підмножинах даних, залежно від того, які змінні мають відсутні значення. Хоча попарне видалення зберігає більше даних, ніж спискове, воно також може призвести до проблем, таких як матриці інтеркореляції, які не є позитивно ви-

значеними, що потенційно перешкоджає подальшому аналізу.

Статистична імпутація для потоків передбачає адаптацію традиційних статистичних методів, таких як імпутація середнього або медіанного значення до контексту потоку, часто за допомогою ковзних вікон. У ковзному вікні останніх точок даних обчислюють середнє значення або медіану спостережуваних значень для певного атрибута та використовують для імпутації будь-яких відсутніх значень у цьому вікні. Цей підхід є обчислювально простим і може бути ефективним, коли механізмом відсутніх даних є MCAR або MAR, а розподіл даних у вікні є відносно стабільним [7].

Імпутація потоків на основі машинного навчання набула поширеності завдяки своїй здатності фіксувати складніші взаємозв'язки та адаптуватися до змін у розподілі даних. Алгоритми онлайн-навчання, такі як онлайн-градієнтний спуск та адаптивна фільтрація, здатні постійно оновлювати моделі імпутації в процесі надходження нових даних, що дозволяє їм ефективно адаптуватися до концептуального дрейфу в режимі реального часу. Рекурентні нейронні мережі (RNN), зокрема LSTM, були також предметом досліджень завдяки своїй здатності моделювати часові залежності в потокових даних, що забезпечує точне імпутування відсутніх значень [8].

У розподілених середовищах методи розподіленого доповнення спрямовано на обролення відсутніх даних без необхідності об'єднання всіх даних у централізоване сховище, що може бути нездійсненним через проблеми конфіденційності чи обмеження мережі. Такі методи, як ефективна для комунікації розподілена множинна імпутація, були розроблені для горизонтально розподілених даних, де дані з різних джерел або сайтів містять однакові атрибути, але для різних суб'єктів.

Інші спеціалізовані підходи охоплюють імпутацію на основі кореляції, яка використовує зв'язки між різними потоками даних або датчиками для визначення відсутніх значень в одному потоці на основі спостережуваних значень у корельованих потоках. Адаптивна імпутація на основі нечіткої логіки використовує нечітку логіку для управління невизначеністю, пов'язаною з відсутніми значеннями, та адаптації процесу імпутації на основі характеристик незбалансованих потоків даних. Методи доповнення на основі графів будують графіки подібності екземплярів даних у певному часовому вікні, а потім використовують методи поширення повідомлень на цих графіках для використання кореляцій між екземплярами та імпутації відсутніх значень [9].

Наявні методології оброблення відсутніх даних у потоках демонструють чітку еволюцію від простих статистичних методів до більш складних підходів машинного навчання та розподілених обчислень. Спільною характеристикою багатьох досліджень із зазначеним питань є визнання часової та розподіленої природи даних, а також потреба в методах, які можуть адаптуватися до змін у розподілі базових даних.

Гібридний підхід до доповнення потокових даних у розподілених системах

Щоб усунути обмеження наявних методів оброблення неповних даних у розподілених потоках для інтелектуального аналізу, пропонується новий підхід, який використовує розподілену природу даних, охоплює контекстну інформацію, адаптується до динамічних характеристик і підтримує обчислювальну ефективність. Цей метод спирається на принципи спільного навчання, контекстно-залежного моделювання та адаптації концептуального дрейфу.

Запропонований підхід може використовувати спільне навчання або федеративні навчальні фреймворки, де кілька розподілених вузлів сприяють побудові глобальної моделі імпутації без необхідності обміну необробленими, потенційно конфіденційними даними.

Федеративне навчання – це метод машинного навчання, за якого моделі навчаються на розподілених даних, що зберігаються на пристроях або серверах, без необхідності централізованого збору чи передачі цих даних. Замість того, щоб передавати всі дані на сервер для навчання, лише параме-

три чи оновлення моделі обмінюються між пристроями та сервером, що дозволяє зберегти конфіденційність даних, оскільки вони не залишають своїх локальних джерел. У цій парадигмі кожен розподілений вузол навчатиме локальну модель імпутації, використовуючи власний потік даних. Потім центральний сервер агрегуватиме ці локально навчені моделі, потенційно використовуючи такі методи, як федеративне усереднення, для створення більш надійної та узагальненої глобальної моделі імпутації. Ця стратегія не лише вирішує проблеми конфіденційності, зберігаючи необроблені дані децентралізованими, а й використовує колективні знання, вбудовані в дані з різних розподілених джерел [10].

Крім того, запропонований метод має зосереджуватися на включенні контекстуальної інформації, щодо потоків даних. Це може охоплювати врахування часових залежностей у кожному потоці, просторових зв'язків між точками даних (якщо дані мають просторовий компонент, наприклад, у сенсорних мережах) та інших відповідних контекстуальних факторів. Приміром, у випадку даних датчиків, процес імпутації може враховувати показники сусідніх датчиків у певні проміжки часу чи історичні закономірності того ж датчика у конкретних умовах.

Враховуючи динамічну природу потоків даних, запропонований метод також має охоплювати механізми адаптації до динамічних характеристик, зокрема, концептуального дрейфу. Це може охоплювати постійний моніторинг продуктивності глобальної моделі імпутації вхідних потоків даних на кожному розподіленому вузлі. Якщо виявлено значний дрейф у розподілі даних, система може ініціювати перенавчання або оновлення локальних моделей та подальшу повторну агрегацію на центральному сервері. Для підвищення здатності моделі адаптуватися до змінних шаблонів даних можна використовувати такі методи, як рання зупинка на боці клієнта під час локального навчання або адаптивна оптимізація з урахуванням дрейфу на боці сервера під час агрегації.

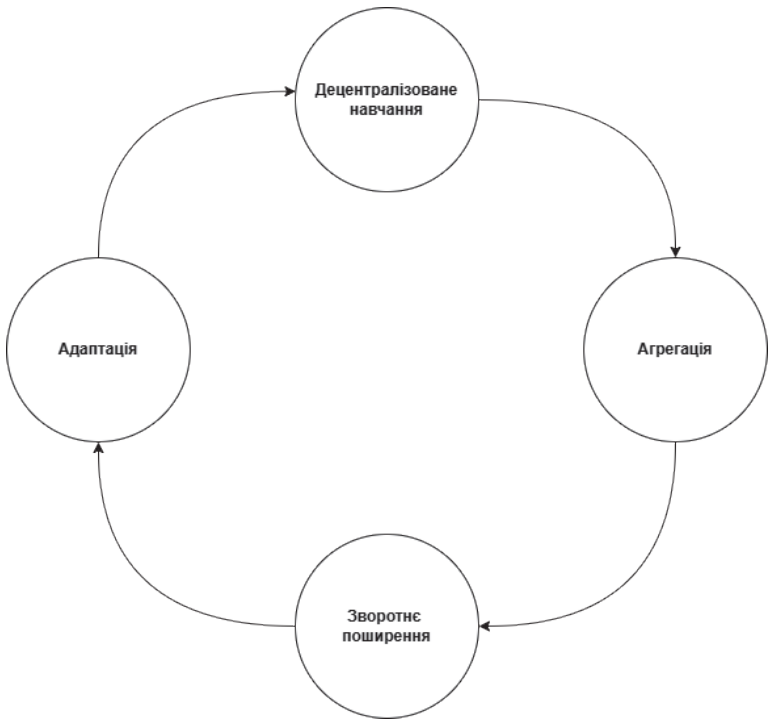

Рис. 1. Етапи пропонованого гібридного підходу до оброблення неповних потокових даних

Для забезпечення доцільності використання гібридного підходу у розподілених середовищах з обмеженими ресурсами, локальні моделі імпутації, навчені на окремих вузлах, мають бути обчислювально ефективними. Такі методи, як дистиляція

знань, коли знання складної моделі переносяться на меншу, ефективнішу модель, можна використовувати для створення моделей імпутації, які здатні ефективно працювати на периферійних пристроях без надмірного споживання ресурсів.

Пропонований гібридний підхід до оброблення неповних потокових даних у розподілених системах реального часу складається з чотирьох етапів та може бути поданий як циклічний процес (рис. 1).

На першому етапі децентралізованого навчання кожен розподілений вузол ініціалізує та навчає спрощену локальну модель імпутації на своєму конкретному потоці даних, зокрема, відповідну контекстуальну інформацію. На етапі агрегації параметрів центральний сервер керує процесом федеративного навчання, де він періодично агрегує параметри й оновлення з локальних моделей імпутації, навчених на кожному вузлі. Іншим етапом є зворотнє поширення, на якому отримана глобальна модель імпутації набуває необхідних характеристик, використовуючи переваги колективного навчання на всіх вузлах, а потім розподіляється у звортньому напрямку на окремі вузли. Кожен вузол використовує глобальну модель для імпутування відсутніх значень у свій вхідний потік даних у режимі реального часу. Цю глобальну модель можна потенційно додатково налаштувати, використовуючи локальні дані на кожному вузлі, щоб врахувати будь-які унікальні локальні характеристики. Останнім етапом є процес адаптації, що виконується періодично, під час якого система постійно контролює продуктивність процесу імпутації на кожному вузлі на наявність ознак концептуального дрейфу. Виявивши значний дрейф, система ініціює новий раунд локального навчання та глобальної агрегації для оновлення моделей імпутації.

Математично можемо визначити підхід таким чином: нехай $N = \{ N_1, N_2 \dots N_n \}$ – множина розподілених вузлів, $D_t^i = (x_j^t, m_j^t)$ – набір даних на вузлі N_i в момент часу *t,* де $x_j^t \in \mathbb{R}$ – *j*-та точка даних у момент часу t, $m_j^t \in \{0, 1\}$ – вектор бінарної маски відсутності, f_i^t – локальна модель імпутації на вузлі N_i в час *t*, $\theta_i^t -$ параметри локальної моделі f_i^t, тоді для кожного етапу можемо записати наступні твердження.

Кожен вузол знаходить вектор відсутніх даних:

$$\hat{x}_j^t = f_i^t \left(x_j^t, m_j^t, c_j^t\right)$$

де c_j^t – є контекстною інформацією (наприклад, залежності часових рядів, показники сусідніх датчиків, семантичні вбудовування для табличних даних). Локальна модель навчена мінімізувати втрати від маскованої реконструкції:

$$L_i^t = \sum_{j=1}^{D_t^i} |(1 - m_j^t) \odot (\hat{x}_j^t - x_j^t)|^2$$

Через задані проміжки часу вузли надсилають оновлення параметрів моделі θ_i^t на центральний сервер:

$$\theta^{t+1} = \sum_{i=1}^{n} \frac{w_i^t}{\sum_k w_k^t} \theta_i^t$$

де w_i^t – ваговий коефіцієнт, наприклад, пропорційний кількості зразків або оцінці продуктивності.

На етапі адаптації кожен вузол оцінює достовірність прогнозування або розподіл помилок з плином часу. Якщо виявляється дрейф (наприклад, за допомогою методу виявлення змін, такого як Пейдж-Хінклі), це запускає локальне перенавчання. За методом Пейдж-Хінклі, нехай ε_i^t – середнє ковзне помилки імпутації, тоді при:

$$\left|\varepsilon_i^t - \varepsilon_i^{t-k}\right| > \delta$$

де k – кількість кроків зворотнього порівняння, δ – межа чутливості помилки, вузол перенавчає свою модель та починає новий раунд федеративних оновлень [12].

Моделювання роботи гібридного підходу

Для оцінки нового підходу проведемо моделювання роботи запропонова-

ного алгоритму з використанням засобів мови програмування Golang. Порівняльну характеристику метрик щодо моделювання роботи методів імпутації подано у табл.1.

Таблиця 1

Порівняльна характеристика метрик щодо моделювання роботи методів імпутації

Метод	RMSE	MAE	Bandwith load
Гібридний підхід	6.2	3.8	21.5 мб
Mean Imputation	12.5	9.1	-
Deep Autoencoder	7.1	5.2	1974 мб

Як тестовий набір даних було використано набір «ElectricityLoadDiagrams20112014» (https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014), що репрезентує енергоспоживання 370 домогосподарств з оновленням даних один раз на 15 хвилин протягом 2011-2014 років. Цей датасет моделює мережу з розподілених датчиків, що надсилають дані на централізований вузол оброблення. І на прикладі сезонності даних – зростання або спадання споживання електроенергії залежно від пори року можна протестувати роботу запропонованого методу з концептуальним дрейфом.

Для моделювання дані розподіляють між змодельованими вузлами, водночас 20 значень випадковим чином видаляють за допомогою схем MAR і MNAR. Для порівняння використаємо методи доповнення середнім (Mean Imputation) і метод глибокого автоенкодера (Deep Autoencoder). Оцінку ефективності проведемо за допомогою метрики середньоквадратичного відхилення (RMSE) та середнього абсолютного відхилення (MAE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2};$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|$$

Додатково обчислимо загальну кількість даних обміну між вузлами під час роботи для визначення навантаження на мережу (Bandwidth load). Для навчання моделей використаємо бібліотеку golearn. Модель розподіленої системи складається з 370 вузлів.

Відповідно до отриманих результатів моделювання (табл. 1) запропоновано гібридний підхід до доповнення даних, що показав кращий результат за метрикою RMSE порівняно з методом Deep Autoencoder на 13%, водночас зменшивши середньоквадратичне відхилення порівняно з методом Mean Imputation удвічі. За метрикою MAE приріст точності порівняно з Deep Autoencoder склав 27%.

Водночас гібридний підхід значно зменшив навантаження на мережу з 1974 мегабайт даних до 21.5 мегабайт порівняно з методом Deep Autoencoder. Це пояснюється зниженням розміру пакетів, що передаються між вузлами через тип даних. Якщо Deep Autoencoder вимагає передачу сирих даних з вузлів для обчислення відсутніх даних, то в пропонованому підході між вузлами передаються лише параметри локальної і агрегованої моделі, що значно впливає на обсяг споживання мережевого трафіку та пропускної здатності мережі.

Запропонований метод, хоча й перспективний, однак має потенційні обмеження. Накладні видатки на зв'язок, пов'язані з процесом федеративного навчання, потребують ретельного керування, особливо в середовищах з обмеженою пропускною здатністю або високою затримкою. Оброблення значної неоднорідності в розподілі даних між різними вузлами також може створювати проблеми, вимагаючи складних методів агрегації. Складність точного виявлення та адаптації до різних типів концептуального дрейфу в розподіленому середовищі вимагає подальшого дослідження. Крім того, є компроміс між складністю та точністю моделей імпутації та їхньою обчислювальною ефективністю, який необхідно ретельно збалансувати на основі вимог конкретного застосування. Водночас оцінка ефективності методу ім-

путації в контексті конкретних задач інтелектуального аналізу, що виконуються над доповненими потоками даних, має вирішальне значення для підтвердження його ефективності.

Висновки

Проблеми, що виникають через неповні дані в розподілених потоках для інтелектуального аналізу, є значними та поширеними в різних прикладних галузях. Традиційні методи імпутації даних, часто розроблені для статичних наборів даних, намагаються ефективно враховувати динамічний характер, обчислювальні обмеження та вимоги до реального часу цих потокових середовищ.

Запропонований метод імпутації на основі федеративного навчання визначає перспективний напрямок, дозволяючи розподіленим вузлам спільно створювати глобальну модель імпутації. Метод вирішує проблеми конфіденційності та використовує колективний інтелект. Включення контекстної інформації та постійна адаптація до концептуального дрейфу спрямовані на підвищення точності та надійності імпутованих даних з часом.

Хоча вищезгаданий метод пропонує кілька потенційних переваг, такі обмеження, як накладні витрати на зв'язок, обробка неоднорідності даних та складність виявлення та адаптації дрейфу, потребують ретельного розгляду та врахування в майбутніх дослідженнях.

Зазначимо, що ефективні методи доповнення даних мають першочергове значення для розкриття повного потенціалу інтелектуального аналізу потокових даних у РСРЧ. Зі зростанням обсягу та швидкості цих потоків потреба в інноваційних та адаптивних рішеннях для оброблення неповних даних ставатиме критичнішою. Запропонований спільний, контекстно-залежний та адаптивний метод імпутації є позитивним кроком до вирішення цих проблем і забезпечення надійнішого та ефективнішого аналізу великих обсягів даних, що генеруються в розподілених середовищах.

References

1. Handling missing values in data streams: An overview. Afonso Lima, Elaine P. M. de Sousa, 2024.
2. Distributed Data and Immersive Collaboration. Daniel Reed, Roscoe Giles, Charles E. Catlett, 1997.
3. Missing Data Imputation: A Comprehensive Review. Journal of Computer and Communications. Alwateer, M. , Atlam, E. , El-Raouf, M. , Ghoneim, O. and Gad, I., 2024.
4. A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms. Youran Zhou, Sunil Aryal, 2024.
5. REAL-TIME SYSTEMS Design Principles for Distributed Embedded Applications. Hermann Kopetz, 1997.
6. Efficient Join Processing Over Incomplete Data Streams (Technical Report). Weilong Ren, Xiang Lian, Kambiz Ghazinour, 2019.
7. Emerging Issues in Data Storytelling CHAPTER 1 | The Challenges of Working With Incomplete Data Sets [Online] – Available from: https://www.icom.org/publications/data-storytelling/the-challenges-of-working-with-incomplete-data-sets (Accessed 28.04.2025).
8. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. GARY KING, JAMES HONAKER, ANNE JOSEPH, KENNETH SCHEVE, 2001.
9. Chukwuemeka Obasi, Victor Oisamoje, Braimoh Ikharo. Security in Distributed System: A Review Perspective, 2022.
10. Time-Sensitive Networking [Online] – Available from: https://campaign.advantech.online/en/global/solutions/intelligent-transportation-systems/resources/white-papers/Time-Sensitive-Networking.pdf , (Accessed 28.04.2025).
11. Concept Drift [Online] – Available from: https://www.iguazio.com/glossary/concept-drift/ , (Accessed 28.04.2025).
12. Continuous Inspection Schemes. Biometrika 41. E. S. Page. 1954.

Одержано: 01.05.2025
Внутрішня рецензія отримана:17.05.2025
Зовнішня рецензія отримана: 18.05.2025

Про авторів :

Плескач Валентина,
д.е.н., к.т.н., професор
https://orcid.org/0000-0003-0552-0972

Жилюк Ярослав,
аспірант
https://orcid.org/0009-0008-9341-9164

Місце роботи авторів:

Київський національний університет
імені Т.Г. Шевченка, факультет
інформаційних технологій
v.pleskach64@gmail.com