

**МОДИФІКАЦІЯ МЕТОДУ DDM
ДЛЯ ВИРІШЕННЯ ЗАДАЧ ПОТОКОЇ ОБРОБКИ ДАНИХ
З ВИКОРИСТАННЯМ СІМЕЙСТВА МОДЕЛЕЙ eTS**

Працибуда В. В., Петров К. Е.

e-mail: vadym.pratsybuda@nure.ua, kostiantyn.petrov@nure.ua

Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

In real-time machine learning systems, performance may degrade as environmental properties change. Therefore, when working with large data streams, there is a need to develop self-maintaining and self-learning systems. The purpose of incremental machine learning models is to solve this problem. Incremental models can learn and predict on-the-fly. Evolving fuzzy systems can operate in an online mode with the ability to adapt model architecture and parameter values. The evolving Takagi-Sugeno (eTS) family can provide accurate results for data streams. This work presents several methods to handle changes in the environment for eTS.

Прогнозування майбутніх станів середовища є необхідною складовою процесу прийняття рішень, тому використовується у різних галузях: економіці, енергетиці, медицині, логістиці та багатьох інших.

Навколишнє середовище може зазнати певних змін з плином часу, що суттєво впливає на функціонування вже розроблених систем машинного навчання. Для уникнення погіршення функціональних якостей системи зазвичай використовують моделі, які б були здатні одночасно навчатися та видавати результати у режимі онлайн. Такі моделі називають моделями із інкрементальним навчанням. Зміни в середовищі функціонування також призводять до змін властивостей вхідних та вихідних даних для моделей машинного навчання.

Є два основних випадки для змін концепцій: різкі зміни (shift) та продовжні зміни у концепції із часом (drift).

Серед моделей з інкрементальним навчанням можна виділити еволюційні нечіткі моделі, які здатні швидко навчатися і видавати результат зі збереженням достатньо високого значення показника точності. Серед сімейств нечітких моделей можна виділити evolving Takagi-Sugeno (eTS), які здатні працювати із поточковими даними та оновлюватись «на льоту» без необхідності у великій вибірці для навчання.

Базовий алгоритм для онлайн eTS складається із декількох етапів.

На першому етапі алгоритм використовує метод eClustering. Починаючи з другої доданої точки в eClustering розпочинається розрахунок потенціалу щодо попередніх точок. Після розрахунку потенціалу точки відбувається оновлення потенціалів центрів кластерів. Далі потенціал точки порівнюється із оновленими потенціалами центрів

кластерів. Якщо потенціал точки перевищує максимальний потенціал центрів кластерів, то така точка є більш інформативною для даних. Якщо вона знаходиться близько до існуючого кластеру, то стає новим центром цього кластеру. Якщо ж точка розташована за межами радіуса дії кластеру, створюється новий кластер.

На другому етапі, отримавши набір правил (кластерів), модель eTS визначає значення функції належності точки до кластерів за допомогою рекурсивної формули Коші. Отримані значення функцій належності передаються у вихідний лінійний шар.

Після отримання вихідних даних (результатів виведення) здійснюється навчання моделі. Модель eTS може імплементувати локальне та глобальне навчання, використовуючи методи рекурсивного нечітко-зваженого методу найменших квадратів (Fuzzy Weighted Least Squares – FWLS) та рекурсивного методу найменших квадратів відповідно [1].

У випадку концептуальних змін, модель eTS може створити нове правило, яке буде описувати нові реалії. З часом додавання нових правил призводить до значного збільшення розмірності моделі, якого можна уникнути, використовуючи методи очищення та поєднання правил.

Для якісної адаптації розмірності моделі під час концептуальних змін постає питання щодо визначення моменту початку змін.

Для визначення моменту змін, можна використовувати концепт віку кластера [2]. Вік кластера можна визначити через додавання значень функцій належності точок з врахуванням кроку життя правила. Градієнт та друга похідна від віку кластера є змінними, які допомагають визначити момент зміни середовища. У випадку якщо зміни кутового коефіцієнта та середнього значення суттєві, то друга похідна віку вказує на аномальні точки, які належать новому стану середовища.

У випадку eTS, можна використовувати інформацію про матриці Гессе або коваріації. У коваріаційній матриці на головній діагоналі розташовані значення дисперсії. Зі збільшенням кількості прикладів зі статичного середовища, значення елементів головної діагоналі не мають зростати з часом, оскільки модель стає «впевненішою» у своїх прогнозах. Найпростішим способом відслідковування значень елементів головної діагоналі є слід матриці. Збільшення значення суми дисперсії зумовлене змінами у середовищі, а отже це означає, що відбуваються концептуальні зміни. Для визначення локальної нестабільності в правилах можна зберігати значення дисперсій кластерів за певного значення вікна.

Найпростішим варіантом визначення змін є використання методу контрольованого Drift Detection Method (DDM) [3]. За допомогою методу обчислюють мінімальну похибку та стандартне відхилення похибки моделі за час роботи. Результати, отримані на виході моделі, не повинні погіршуватись з часом у середовищі без змін. Тому з врахуванням значення дисперсії має виконуватися правило трьох сигм: вихід дисперсії

похибки та середньої похибки за межі, визначені правилом, має сигналізувати про зміну середовища, що повинна спонукати модель до змін у параметрах та архітектурі. Метод «знає» про мітки, в яких відбуваються зміни, проте «не знає» про правила, які впливають на ці мітки, оскільки метод працює з глобальними даними про похибки, а не з локальними кластерами.

В роботі пропонується модифікація DDM для роботи із локальними даними кластерів. Значення похибки має обчислюватися за допомогою відносної похибки щоб уникнути великих значень помилок у задачах регресії. Для роботи з локальними даними використовується зважування похибки за допомогою функцій належності точки до правил. У результаті отримується значення локальної похибки правила. Потрібно розрахувати середнє значення локальною похибки для кожного правила на основі минулих значень середнього. Варто встановити порогове значення, при якому середнє значення похибки правила буде оновлюватись, оскільки нечіткі моделі виконують м'які розрахунки, а отже правила породжують низькі значення функції належності навіть для віддалених точок. Після розрахунку середньої похибки варто розрахувати середнє значення дисперсії. Отримавши дисперсію, можна виконати перевірку за правилом трьох сигм. Якщо правило трьох сигм не виконується, то локальні похибки вказують на кластер, на який вплинули зміни середовища. У цьому випадку варто збільшити ваги, що призначаються новим вхідним точкам. Для захисту від викидів у вхідних даних рішення про концептуальні зміни слід приймати лише у випадку їх спостереження протягом певного часу.

Оскільки методи, що базуються на використанні віку кластера та коваріаційної матриці, є неконтрольованими, то не отримують інформацію про вихід моделі. Тому такі методи не будуть працювати при зміні концепцій у вихідних даних за стабільних вхідних.

Запропонована модифікація DDM є контрольованою, а тому має інформацію про похибки у виході моделі та здатна реагувати на зміни між розподілами вхідних і вихідних даних. Розроблена модель здатна швидко реагувати на зміни, проте має вищі обчислювальну та просторову складності порівняно з моделлю на основі віку.

Список використаних джерел:

1. Lughofer E. Evolving Fuzzy Systems – Methodologies, Advanced Concepts and Applications. Berlin, Heidelberg : Springer Berlin Heidelberg, 2011. DOI: <https://doi.org/10.1007/978-3-642-18087-3>.
2. Kacprzyk J., Lin T. Y., Liau C.-J. Granular, Fuzzy, and Soft Computing. Springer, 2023.
3. Learning with Drift Detection / J. Gama та ін. Advances in Artificial Intelligence – SBIA 2004. Berlin, Heidelberg, 2004. С. 286–295. DOI: https://doi.org/10.1007/978-3-540-28645-5_29.