Issue 16, 2024

https://doi.org/10.23939/sisn2024.16.133

УДК 004.9

ІНФОРМАЦІЙНА СИСТЕМА НАПІВКОНТРОЛЬОВАНОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ВИБІРОК ДАНИХ ІЗ ВИСОКОЮ РОЗМІРНІСТЮ

Неля Мірошниченко[1], Ірина Перова[2], Олег Дацок[3]

[1-3] Харківський національний університет радіоелектроніки,
кафедра системотехніки, Харків, Україна
[1] E-mail: nelia.miroshnychenko@nure.ua, ORCID: 0000-0002-3846-1668
[2] E-mail: iryna.perova@nure.ua, ORCID: 0000-0003-2089-5609
[3] E-mail: oleh.datsok@nure.ua, ORCID: 0000-0003-4489-3819

Дослідження великих обсягів даних для виявлення прихованих закономірностей і тенденцій стає дедалі важливішим і кориснішим у останні роки. Ці великі обсяги даних характеризуються широкою доступністю, складністю структур і значним обсягом інформації.

У статті запропоновано детальний опис інформаційної системи напівконтрольованого навчання для аналізу вибірок даних із високою розмірністю. Систему розроблено з метою опрацювання великих обсягів даних і використано методи напівконтрольованого навчання для ефективного аналізу та класифікації. Для цього проаналізовано наявні інформаційні системи, які здатні працювати з вибірками даних, які мають високу розмірність, а також методи, що ефективно проводять аналіз та класифікацію цих вибірок даних. Детально описано архітектуру системи, зокрема методи опрацювання даних, вибір ознак, модулі передпроцесінгу та методи оптимізації навчання.

Ключові слова: інформаційна система напівконтрольованого навчання, вилучення ознак, великі вибірки даних, груповий пошук ознак, нечітка адаптивна мережа Кохонена.

Вступ

Останнім часом аналіз високорозмірних наборів даних для виявлення закономірностей та класифікації стає все більш важливим у різних галузях, зокрема машинне навчання, науку про дані та медицину. Зростання обсягу та складності даних потребує використання високоефективних методів опрацювання та отримання розуміння з цих наборів даних [1].

Доступність та обсяг великих наборів даних стрімко зростають завдяки прогресивним технологіям збору даних та поширенню цифрових платформ. Ці набори даних часто відображають складні патерни та кореляції, що робить їх цінними джерелами інформації для прийняття рішень та прогнозування [2]. Тож необхідно застосовувати інформаційні системи, які будуть здатні працювати з цими даними. Наприклад, традиційні методи навчання з учителем мають обмеження при роботі з високорозмірними даними через проблеми, як-от рідкість даних, зайві ознаки та прокляття розмірності. Проте інформаційні системи напівконтрольованого типу при навчанні системи використовують як розмічені, так і нерозмічені дані, що дає можливість відкривати перспективи для вирішення цих проблем та отримувати значущі висновки з складних наборів даних.

Формулювання проблеми

Вибірки даних із високою розмірністю все частіше трапляються при аналізі медичних даних, що створює необхідність у розробці ефективних методів опрацювання та аналізу таких даних. Однією з ключових проблем, яка виникає при роботі з високорозмірними медичними даними, є складність їхньої структури та велика кількість ознак. Це може призводити до проблеми вибору найбільш інформативних ознак та зменшення розмірності даних без втрати важливої інформації. Додатково існує проблема ефективної класифікації цих даних, оскільки традиційні методи машинного навчання можуть бути неефективними через прокляття розмірності та рідкість даних. Тому доцільним є розроблення інформаційних систем, здатних працювати з високорозмірними медичними даними та використовувати напівконтрольовані методи навчання для ефективного аналізу та класифікації цих даних.

Аналіз останніх досліджень та публікацій

Інформаційні системи напівконтрольованого навчання є важливою галуззю в машинному навчанні, яка зосереджується на розробці алгоритмів та методів для навчання моделей з використанням і підтверджених, і непідтверджених даних. Ці системи використовують невелику кількість маркованих (підтверджених) прикладів разом із великою кількістю немаркованих (непідтверджених) даних для покращення точності моделей. В останні роки ця галузь отримала значний інтерес з боку дослідників у зв'язку з можливістю ефективного використання невпорядкованих даних, які часто є більш доступними в реальних застосуваннях. Проаналізувавши останні дослідження та публікації з інформаційних систем напівконтрольованого навчання, було знайдено різноманітні методи, алгоритми та застосування цих систем у практичних завданнях машинного навчання, як-от: розпізнавання образів, класифікація даних та аналіз тексту.

Протягом останніх десятиліть напівконтрольоване навчання в класифікації вважалося одним із найбільш активних дослідницьких галузей через зростаючий фізичний попит. Було розглянуто дослідження, в якому представлено графовий напівконтрольований алгоритм навчання під назвою GSB2LS у межах Байєса для класифікації. Алгоритм ефективно досліджував немічені дані, застосовуючи складений попередній, який складається з неміченої багатоманітної інформації та розріджених байєсівських висновків для широкої структури. Зокрема, GSB2LS використовував переваги широкої структури для пошуку більшої кількості потенційних асоціацій ознак, різноманітну регуляризацію для охоплення вигідної взаємозалежності немічених зразків, байєсівську структуру для підтримки універсальної розрідженості, швидку максимізацію граничної ймовірності для оновлення набору релевантності на основі на визначений внесок, що призводив до можливості опрацювання великомасштабних даних індуктивним способом. Окрім того, алгоритм здатний виводити ймовірнісну оцінку прогнозу для подальшого аналізу рішень. Великі емпіричні результати підтвердили чудову продуктивність алгоритму [3].

Розглянуто застосування напівконтрольованого машинного навчання на даних метилювання для підвищення точності моделей контрольованого навчання при класифікації пухлин ЦНС. Було застосовано комбінований підхід, який передбачав самонавчання з правленням за допомогою моделі опорних векторних машин (SETRED-SVM) і модель L2-штрафної мультиноміальної логістичної регресії для отримання міток високої надійності з кількох позначених екземплярів. Результати восьми моделей випадкового лісу та нейронної мережі показали, що псевдомітки, отримані за допомогою методу напівконтрольованого навчання, можуть значно підвищити точність прогнозування. Ця комбінація напівконтрольованої техніки та мультиноміальної логістичної регресії має потенціал для ефективного використання великої кількості загальнодоступних немічених даних метилювання. Такий підхід є дуже корисним у наданні додаткових навчальних прикладів, особливо для рідкісних типів пухлин, щоб підвищити точність прогнозування контрольованих моделей [4].

Інформаційні системи напівконтрольованого машиного навчання використовують як інструменти виділення специфічних функцій для проблеми класифікації афективних станів за допомогою

сигналів ЕЕГ. Останнім часом було показано, що фізіологічні сигнали, зокрема електроенцефалограма (ЕЕГ), дуже ефективні в оцінюванні афективних станів користувача під час соціальної взаємодії або під час відео- чи аудіостимулів. Однак через велику кількість параметрів, пов'язаних із нейронним вираженням емоцій, досі залишається багато невідомих щодо конкретної просторової та спектральної кореляції сигналу ЕЕГ та вираження афективних станів. Щоб дослідити таку кореляцію, два типи напівконтрольованих підходів до глибокого навчання, стекований автокодер із шумоподавленням (SDAE) і мережі глибоких вірувань (DBN) були використані в інформаційній системі, яку розглядали. Щоб оцінити ефективність пропонованих напівконтрольованих підходів, було проведено експеримент із класифікації афективних станів для суб'єкта в базі даних DEAP для класифікації двовимірних афективних станів. Модель, заснована на DBN, досягла усереднених балів F1 86,67 %, 86,60 % і 86,69 % для класифікації станів збудження, валентності та симпатії відповідно, що значно покращило сучасну класифікацію. Досліджуючи вектори ваги на кожному прошарку, ми також змогли отримати уявлення про просторове або спектральне розташування найбільш розрізнювальних ознак. Головною перевагою застосування методів напівконтрольованого навчання можна виділити той факт, що в дослідженні використано лише невелику частину позначених даних. Як приклад, саме для цього дослідження використано 1/6 навчальних вибірок [5].

Формулювання цілі статті

Мета статті – розробити інформаційну систему напівконтрольованого навчання для роботи з медичними вибірками даних із високою розмірністю, що дасть змогу працювати з розміченими та нерозміченими даними та ефективно аналізувати і класифікувати ці дані.

Виклад основного матеріалу

Інформаційну систему напівконтрольованого навчання побудовано з трьох модулів, кожен з яких відповідає різним аспектам вибору ознак і зменшенню розмірності даних. Перший модуль відповідає за аналіз вихідних даних та визначення їхньої структури й основних характеристик. Другий модуль використовує різноманітні методи відбору ознак та методу зменшення розмірності, щоб виділити найбільш важливі характеристики даних. Третій модуль виконує важливу роль у фінальній обробці та оптимізації даних після вибору найбільш важливих ознак та зменшення їхньої розмірності, які легше й ефективніше можуть бути оброблені та використані в подальших аналітичних задачах.

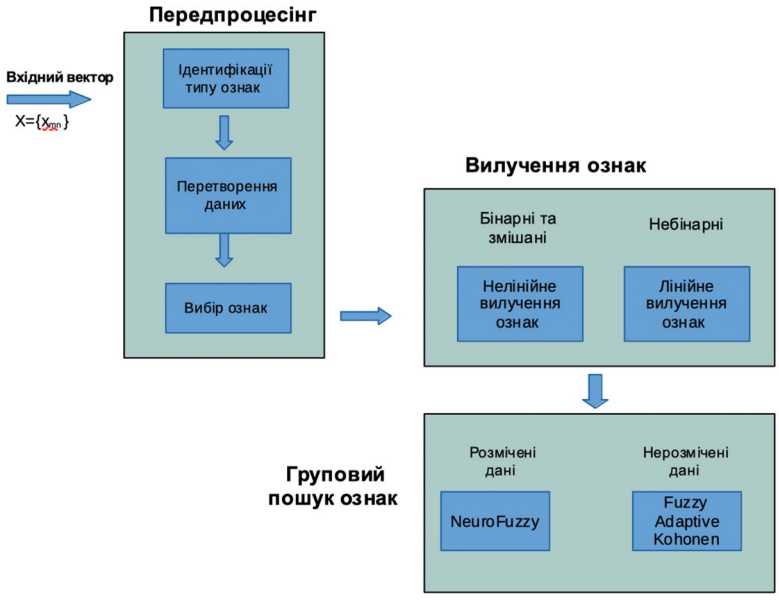

Рис. 1. Інформаційна система напівконтрольованого навчання

На вхід інформаційної системи напівконтрольованого навчання подається набір даних, який складається з переліку пацієнтів та ознак, що їх характеризують. Припустімо, що цей набір даних є матрицею X={x_{mn}} розмірністю K´ k, де K – кількість пацієнтів, а k – кількість ознак, які характеризують кожного пацієнта. Тоді кожен пацієнт описують (k´ 1)-вектором ознак:

$$x(m) = (x_1(m), x_2(m), \dots x_j(m), \dots x_k(m))^T, \quad (1)$$

де m=1,…,K.

Далі вхідний вектор ознак переходить до розробленого модуля передпроцесінгу даних для оптимізації опрацювання вхідного вектору перед вилученням ознак. Рух вхідного вектору через модуль передпроцесінгу складається з кількох етапів, зокрема блоки ідентифікації типу ознак, перетворення даних та вибору ознак.

Вхідний вектор даних спрямовується на перший блок, ідентифікації типу ознак, модуля передпроцесінгу для визначення типу кожної ознаки. Цей блок містить низку операцій і алгоритмів, спрямованих на аналіз характеристик ознак для їхньої класифікації за типом, а саме:

1) аналіз діапазону значень (для бінарних ознак діапазон значень обмежується двома унікальними значеннями (наприклад, 0 або 1). Числові ознаки мають ширший діапазон значень, тоді як категоріальні ознаки мають обмежений набір категорій або міток);

2) визначення унікальних значень (бінарні ознаки мають два унікальних значення, тоді як числові та категоріальні ознаки можуть мати більше унікальних значень);

3) визначення типу ознак (бінарна, числова чи категоріальна);

4) створення міток типу ознак (після визначення типів ознак генеруються мітки або індикатори для подальшого використання у модулях передпроцесінгу) [6].

У результаті проходження вхідного вектора через блок ідентифікації типу ознак кожній означеній ознаці $x_j(m)$ вектора $x(m)$ присвоєно тип: бінарна (binary), числова (numeric) або категоріальна (categorical) ознака.

Після ідентифікації типу ознак вектор даних направляють до блоку перетворення даних. Цей блок містить низку операцій з опрацювання даних, спрямованих на оптимізацію та стандартизацію даних залежно від їхнього типу, а саме:

1) нормалізація числових даних (перетворення значень числових ознак, таким чином, щоб вони мали стандартний діапазон (наприклад, від 0 до 1 або з використанням стандартного розподілу середнє значення 0 і стандартне відхилення 1));

2) кодування категоріальних даних (перетворення категоріальних ознак у числовий формат за допомогою методів, таких як one-hot encoding або використання кодування міток (label encoding), роблячи їх придатними для моделювання);

3) опрацювання пропущених значень (заповнення середнім значенням, медіаною або видалення рядків з пропущеними значеннями) [7].

Ці операції у блоці перетворення даних допомагають стандартизувати та підготувати вхідні дані для переходу у блок вибору ознак, де проводиться відбір найбільш інформативних та релевантних ознак для подальшого використання [8]. Цей блок допомагає зменшити розмірність даних, водночас зберігаючи найважливішу інформацію, що сприяє покращенню ефективності моделювання і зниженню його складності. Для цього застосовуються методи машинного навчання, статистичні методи, різні стратегії (відсік ознак із низькими важливими показниками, використання алгоритмів із вбудованим вибором ознак, методи вибору ознак на основі моделей), а також врахування додаткових критеріїв (взаємна кореляція між ознаками, стабільність важливості під час зміни набору даних, час виконання алгоритму відбору ознак).

На виході з блоку вибору ознак із вхідного вектора ознак формується новий вектор ознак, який враховує всі проведені типи опрацювання і може бути готовим для використання в подальшому аналізі даних. Новий вектор описується як:

$$x_{pre}(m) = (x_1(m), x_2(m), \dots x_j(m), \dots x_k(m))^T. \quad (2)$$

Цей вектор переходить до модуля вилучення ознак. Вказаний модуль містить у собі два блоки – блок вилучення нелінійних ознак та блок вилучення лінійних ознак. Бінарні та змішані дані потрапляють до блоку нелінійного вилучення ознак, а не бінарні – до блоку лінійного вилучення ознак.

Вектор $x_{pre}(m)$, в якому $x_{pre}(m) \in R^{k\times 1}$, де $k \times 1$ – розмірність вхідного вектору ознак, подається на вхід блоку вилучення нелінійних ознак. Першим кроком у роботі блоку є енкодер, в якому відбувається лінійне перетворення $x_{pre}(m)$ у вектор коду $z_{pre}(m)$. Лінійне перетворення можна представити за допомогою матриці ваг W_{enc} та вектора зсуву b_{enc}:

$$z_{pre}(m) = \sigma(W_{enc}x_{pre}(m) + b_{enc}), \quad (3)$$

де σ – нелінійна активаційна функція, яка може бути лінійною (наприклад, ідентична функція) або нелінійною (наприклад, сигмоїдна, гіперболічний тангенс).

Наступним кроком у роботі блоку вилучення нелінійних ознак є робота декодера, в якому відбувається відновлення вихідних даних із кодованого представлення вектора $z_{pre}(m)$ назад у вхідні дані вектора $x'_{pre}(m)$. Згаданий декодер може бути представлений як:

$$x'_{pre}(m) = W_{dec}z_{pre}(m) + b_{dec}\ , \quad (4)$$

де W_{dec} – матриця ваг декодера; b_{dec} – вектор зсуву декодера.

Далі необхідно визначити функцію втрат як середньоквадратичну помилку між вхідними даними, представленими вектором $x_{pre}(m)$ та відновленими даними – $x'_{pre}(m)$ для навчання блоку нелінійного вилучення ознак. Функція втрат представлена як:

$$L = \frac{1}{K}\sum_{m=1}^{K} ||x_{pre}(m) - x'_{pre}(m)||^2\ , \quad (5)$$

де K – загальна кількість пацієнтів.

Функція втрат дає змогу оцінити якість відтворення вхідних даних із кодованого представлення. Для ефективної роботи блоку нелінійного вилучення ознак необхідно мінімізувати функцію втрат L, використати методи оптимізації, як-от стохастичний градієнтний спуск. Необхідно обчислити градієнти функції втрат $L(x_{pre}(m),\ x'_{pre}(m))$ щодо параметрів блоку нелінійного вилучення ознак $W_{enc},\ b_{enc}, W_{dec},\ b_{dec}$. Градієнти обчислюють за допомогою методу зворотнього поширення помилки. Обчислення градієнта функції втрат для декодера таке:

$$\begin{aligned}\nabla_{x'_{pre}(m)}L\left(x_{pre}(m), x'_{pre}(m)\right) &= 2(x'_{pre}(m) - x_{pre}(m))\\ \nabla_{W_{dec}}L\left(x_{pre}(m), x'_{pre}(m)\right) &= \nabla_{x'_{pre}(m)}L\left(x_{pre}(m), x'_{pre}(m)\right)z_{pre}(m)^T\\ \nabla_{b_{dec}}L\left(x_{pre}(m), x'_{pre}(m)\right) &= \nabla_{x'_{pre}(m)}L\left(x_{pre}(m), x'_{pre}(m)\right).\end{aligned} \quad (6)$$

Градієнт функції втрат для енкодера такий:

$$\begin{aligned}\nabla_{z_{pre}(m)}L\left(x_{pre}(m), x'_{pre}(m)\right) &= W_{dec}^T\nabla_{x'_{pre}(m)}L\left(x_{pre}(m), x'_{pre}(m)\right)\\ \nabla_{W_{enc}}L\left(x_{pre}(m), x'_{pre}(m)\right) &= \nabla_{z_{pre}(m)}L\left(x_{pre}(m), x'_{pre}(m)\right)x_{pre}(m)^T\end{aligned} \quad (7)$$

$$\nabla_{b_{enc}} L\left(x_{pre}(m), x'_{pre}(m)\right) = \nabla_{z_{pre}(m)} L\left(x_{pre}(m), x'_{pre}(m)\right).$$

Після обчислення градієнтів функції втрат необхідно оновити параметри блоку нелінійного вилучення ознак у напрямку зменшення функції втрат. Для кожного параметра Q, де Q може бути W_{enc}, b_{enc}, W_{dec}, b_{dec}, оновлення виконується згідно з формулою:

$$\mathrm{Q} \leftarrow Q - \alpha \nabla_Q L\left(x_{pre}(m), x'_{pre}(m)\right), \tag{8}$$

де α – швидкість навчання.

Цей процес повторюється для кожного навчального прикладу $x_{pre}(m)$ у навчальному наборі даних протягом кількох епох навчання. Один прохід через увесь набір даних називають епохою. Через кількість ітерацій або епох функція втрат зазвичай зменшується, що свідчить про навчання моделі.

Після навчання блоку нелінійного вилучення ознак, $z_{pre}(m)$ закодоване представлення може бути використане для задач класифікації. На виході блоку нелінійного вилучення ознак отримуємо вектор $x'_{pre}(m)$, який є реконструйованою версією вхідного вектора $x_{pre}(m)$.

Після того, як вектор пройшов усі блоки модуля вилучення ознак, вектор $x'_{pre}(m)$ переходить до модуля групового пошуку ознак, що складається з блоку NeuroFuzzy та блоку Fuzzy Adaptive Kohonen. У разі, коли вхідний вектор $x'_{pre}(m)$ подається разом із референс вектором y_c, то процес переходить до блоку Neuro Fuzzy, бо дані, які подаються на блок NeuroFuzzy, вже є розміченими.

У вказаному блоці застосовують вагові вектори для розділення даних на класи з урахуванням рівня нечіткості в їхній приналежності до кожного класу [9]. Навчання проводиться ітеративно, оновлюючи ваговий вектор для кожного класу на основі вхідних даних та їхньої приналежності до класу. Це забезпечує ефективну класифікацію навіть тоді, коли дані є неоднозначними.

У роботі блоку NeuroFuzzy передбачено прямі зв'язки, які з'єднують кожен нейрон вхідного рівня з нейронами вихідного рівня, кожному з яких призначено опорний вектор. Блок NeuroFuzzy використовує набір функцій належності для оцінки ступеня асоціації між кожним вхідним вектором і референс векторами. Ця стратегія контролює конкуренцію між референс векторами, даючи змогу кожному вхідному вектору брати участь у процесі навчання і забезпечуючи оптимальні результати класифікації.

Цільова функція, яку необхідно мінімізувати, така:

$$E = \sum_{j=1}^{C} mf_c(x'_{pre}(m))\left\|x'_{pre}(m) - y_c\right\|, \tag{9}$$

де c =1,2…C – набір функцій належностей mf; y_c – референс вектор (еталонний вектор), $x'_{pre}(m)$ – вхідний вектор.

Якщо припустити, що еталонний вектор y_c визначено як переможець, то функцію належності mf_c можна виразити як:

$$mf_c = \frac{1}{1 + \dfrac{\left\|x'_{pre}(m) - y_i\right\|}{\left\|x'_{pre}(m) - y_c\right\|}} \quad \text{if} \quad c \neq 1$$
$$mf_c = 1 \quad \text{if} \quad c = 1. \tag{10}$$

У контрольованому навчанні представлено послідовність тренувальних векторів із відомими категоріями [10]. Розраховується відстань між кожним тренувальним вектором і еталонними векторами для всіх категорій. Оптимальний еталонний вектор неодноразово оновлюється для зменшення відмінності між вихідним результатом NeuroFuzzy та бажаним. Під час процесу навчання можуть виникати дві ситуації. У першому випадку, коли цільовий вектор y_i збігається з тренувальним вектором, референс вектор нейрона-переможця оновлюється за формулою:

$$y_i(t+1) = y_i(t) + \alpha(t)(x'_{pre}(m)(t) - y_i(t))(1 + \sum(1 - mf_{ic})^2). \quad (11)$$

Референс вектор, що не став переможцем, оновлюється за формулою:

$$y_c(t+1) = y_c(t) + \alpha(t)(x'_{pre}(m)(t) - y_c(t))(mf_{ic})^2. \quad (12)$$

У другій ситуації цільовий вектор y_i не збігається з тренувальним вектором, тоді референс вектор нейрона-переможця оновлюється за формулою:

$$y_i(t+1) = y_i(t) - \alpha(t)(x'_{pre}(m)\,(t) - y_i(t))(1 + \sum(1 - mf_{ic})^2). \quad (13)$$

Референс вектор, що не став переможцем, оновлюється за формулою:

$$y_c(t+1) = y_c(t) - \alpha(t)(x'_{pre}(m)(t) - y_c(t))(mf_{ic})^2. \quad (14)$$

Щоб дослідити, наскільки є інтенсивного оновлення еталонних векторів під час процесу навчання, необхідно розрахувати скалярний коефіцієнт підсилення. Він зазвичай зменшується в міру того, як навчальний процес збігається до оптимального стану ($0 < \alpha \leq 1$). Розраховується скалярний коефіцієнт підсилення за формулою:

$$\alpha(t+1) = \left(1 - \frac{t}{N}\right)\alpha(t), \quad \alpha(0) = 0{,}05\,, \quad (15)$$

де N – номер заздалегідь визначеної епохи.

Задля досягнення найкращих результатів роботи нейронної мережі прийнято рішення розглянути скалярний коефіцієнт підсилення як:

$$\alpha(t) = r^{-1}(t)\,, \quad (16)$$

де

$$r(t) = \eta r(t-1) + 1\,, \quad (17)$$

де $0 \leq \eta \leq 1$.

Задля досягнення найбільш оптимального значення цільового вектору y_i необхідно розв'язати систему рівнянь:

$$y_i(t+1) = \begin{cases} y_i(t) + \alpha(t)(x'_{pre}(m)(t) - y_i(t))(1 + \sum(1 - mf_{ic})^2) \\ \text{у разі однакової категорії мережевого виходу та цільового вектора} \\ y_i(t) - \alpha(t)(x'_{pre}(m)(t) - y_i(t))(1 + \sum(1 - mf_{ic})^2) \\ \text{у разі різної категорії мережевого виходу та цільовий вектора} \\ y_i(t) \text{ для референс векторів, які не виграють} \\ mf_c = \dfrac{1}{1 + \dfrac{\|x'_{pre}(m) - y_i\|}{\|x'_{pre}(m) - y_c\|}} \\ \alpha(t) = (\eta r(t-1) + 1)^{-1}, \quad 0 \leq \eta \leq 1. \end{cases} \quad (18)$$

На виході блоку NeuroFuzzy отримуємо індекс або мітку класу, до якого вважається найбільш вірогідним приналежність вхідного вектора $x'_{pre}(m)$. Цей результат може бути використаний для подальшого аналізу та інтерпретації отриманих результатів.

Повернувшись до модуля групового пошуку ознак, отриманий на виході блоку нелінійного вилучення ознак вектор $x'_{pre}(m)$ може перейти до блоку Fuzzy Adaptive Kohonen, якщо вхідний вектор подається на вхід блоку без референс вектора, а це означає, що дані не були попередньо розмічені.

Нечіткі адаптивні мережі кластеризації Кохонена пропонують метод групування даних, який поєднує принципи нечіткої логіки з адаптивними мережами Кохонена. Тут кожен нейрон в адаптивній мережі Кохонена функціонує як центральна точка для кластера, а його ваги вказують на відстані між ними [11]. Примітно, що нейрони можуть бути частиною кількох кластерів одночасно, кожен з яких має різні рівні членства. Під час навчання мережі вагові коефіцієнти нейронів точно налаштовуються, щоб мінімізувати розриви між вхідними даними та центрами кластерів, забезпечуючи оптимізований розподіл кластерів у ландшафті даних.

Критерій мінімізації можна представити так:

$$E(mf_j(m), centr_j) = \sum_{m=1}^{KL}\sum_{j=1}^{T} mf_j^{\gamma}(m) Dist^2\left(x'_{pre}(m), centr_j\right) + \sum_{j=1}^{T} r_j \sum_{m=1}^{KL} 1 - mf_j(m))^{\gamma}, \quad (19)$$

де $mf_j(m)$ – ступінь належності вектора $x'_{pre}(m)$ j-ого кластера, $centr_j$ – центр j-ого кластера, KL – к-ть кластерів, T – загальна к-ть точок даних, γ – невід'ємний параметр (фаззифікатор), $Dist\left(x'_{pre}(m), centr_j\right)$ – відстань між $x'_{pre}(m)$ і $centr_j$, r_j – скалярний параметр, значення якого більше 0, що визначає відстань, на якій рівень належності приймає відповідне значення.

Після критерію мінімізації (19) утворюється система рівнянь:

$$\begin{cases} \frac{\partial E(mf_j(m), centr_j)}{\partial mf_j(m)} = 0, \\ \frac{\partial E(mf_j(m), centr_j)}{\partial r(m)} = 0, \\ \nabla_{centr_j} E\left(mf_j(m), centr_j\right) = \vec{0}. \end{cases} \quad (20)$$

Розв'язавши систему рівнянь (20), отримуємо результат:

$$mf_j^{pos}(m) = \left(1 + \left(\frac{Dist^2\left(x'_{pre}(m), centr_j\right)}{r_j}\right)^{\frac{1}{1-\gamma}}\right)^{-1}, \quad (21)$$

$$r_j = \frac{\sum_{m=1}^{KL} mf_j^{\gamma}(m) Dist^2\left(x'_{pre}(m), centr_j\right)}{\sum_{m=1}^{KL} mf_j^{\gamma}}. \quad (22)$$

Розв'язок 3-го рівняння в системі (20) для евклідової норми

$$Dist^E\left(x'_{pre}(m), centr_j\right) = \left\|x'_{pre}(m) - centr_j\right\| = \sqrt{\left(x'_{pre}(m) - centr_j\right)^T \left(x'_{pre}(m) - centr_j\right)}$$

такий:

$$centr_j^{pos} = \frac{\sum_{m=1}^{KL} mf_j^{\gamma}(m) x'_{pre}(m)}{\sum_{m=1}^{KL} mf_j^{\gamma}(m)}. \quad (23)$$

Розглянувши модуль групового пошуку ознак, який складається з двох блоків, стає зрозуміло, що блок NeuroFuzzy оперує маркованими даними, де об'єкти вже класифіковані за певними категоріями або класами. Цей блок інформаційної системи використовує методи класифікації для автоматичного присвоєння нових об'єктів відомим класам шляхом аналізу їхніх характеристик і шаблонів із позначених даних. Другий блок модуля, Fuzzy Adaptive Kohonen, працює з даними без міток, де класи або категорії об'єктів невідомі заздалегідь. Цей блок використовує методи кластеризації для групування об'єктів на основі їхньої подібності або відстані один від одного. Обидва блоки модуля сприяють вирішенню різноманітних завдань, як-от: аналіз клієнтів, прогнозу-

вання тенденцій або виявлення аномалій у даних. Загальна інформаційна система створює ефективний інструмент для аналізу та розуміння великих обсягів даних і позначених, і непозначених.

Щоб оцінити ефективність інформаційної системи напівконтрольованого навчання було використано медичну вибірку даних, яка містить двійкові, дискретні, категоріальні та порядкові типи даних. Вибірка складається з 300 спостережень із 199 ознаками. Для дослідницьких цілей була використана вибірка з 55 спостережень. Своєю чергою дані мали змішаний характер, а саме: деякі були розмічені, а деякі – ні. Після завершення роботи системи, через модулі та блоки якої пройшла вибірка даних, отримано інформацію про точність моделі, що відображає, наскільки точно модель кл0430сифікувала дані для кожного класу. Ці результати представлено у вигляді матриці помилок для розмічених та окремо – для нерозмічених даних. Для покращення розуміння результатів створено візуалізації (рис. 2).

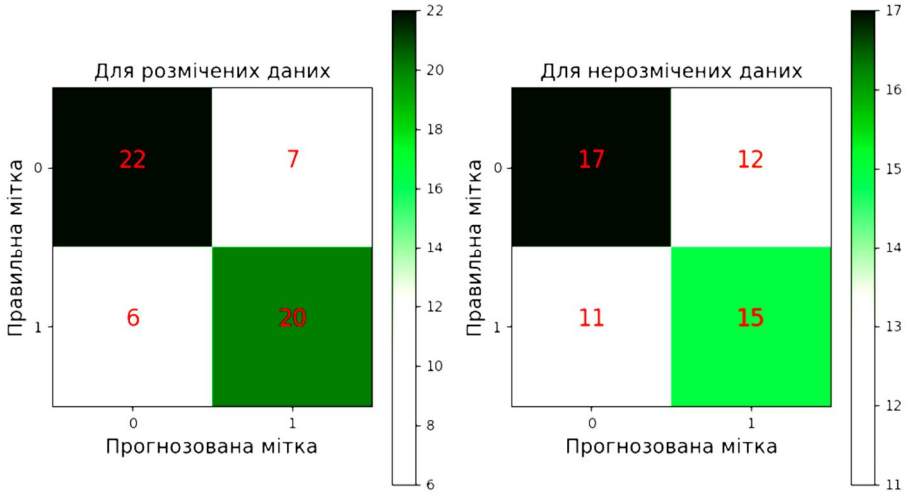

Рис. 2. Матриця помилок для розмічених та нерозмічених даних

З візуалізацій можна побачити, що вказана матриця використовується для визначення кількості правильних та неправильних класифікацій моделі на основі вхідних даних і їхніх відповідних міток. Кожний елемент матриці описує результат класифікації щодо істинної мітки класу.

У верхньому лівому куті матриці знаходиться істино позитивне значення. Воно представляє кількість об’єктів, які насправді належать позитивному класу і були правильно класифіковані моделлю як позитивні.

У верхньому правому куті матриці знаходиться хибно позитивне значення. Воно показує кількість об’єктів, які насправді належать до негативного класу, але були помилково класифіковані моделлю як позитивні.

Хибно негативне значення знаходиться у нижньому лівому куті матриці. Воно вказує на кількість об’єктів, які насправді належать позитивному класу, але були помилково класифіковані моделлю як негативні.

Істино негативне значення знаходиться у нижньому правому куті матриці. Воно відображає кількість об’єктів, які насправді належать негативному класу і були правильно класифіковані моделлю як негативні.

Ці значення допомагають оцінити різні аспекти класифікації моделі і були використані для обчислення різних метрик, як-от: точність, чутливість, специфічність, що визначають продуктивність моделі щодо кожного з класів. Результати аналізу метрик представлено у табл. 1.

Таблиця 1

Порівняльний звіт про класифікацію розмічених та нерозмічених даних

Модель	Точність визначення позитивних класів	Чутливість	F1-показник	Загальна точність моделі
Для розмічених даних	0,76	0,76	0,76	0,75
Для нерозмічених даних	0,58	0,58	0,58	0,58

Проаналізувавши отримані показники продуктивності моделі, стає зрозуміло, що для отримання вищої продуктивності, необхідно і надалі удосконалювати модель. Проте з результатів чітко видно, що метрики для нерозмічених даних мають гірші результати. Це пов'язано з тим, що немає чітко визначених міток класів для порівняння з прогнозами моделі, що ускладнює роботу. Дані висновки підтверджено графіком, представленим на рис. 3, який демонструє тривимірне розташування даних, отриманих за допомогою методу головних компонент (PCA) для двох відомих класів розмічених даних.

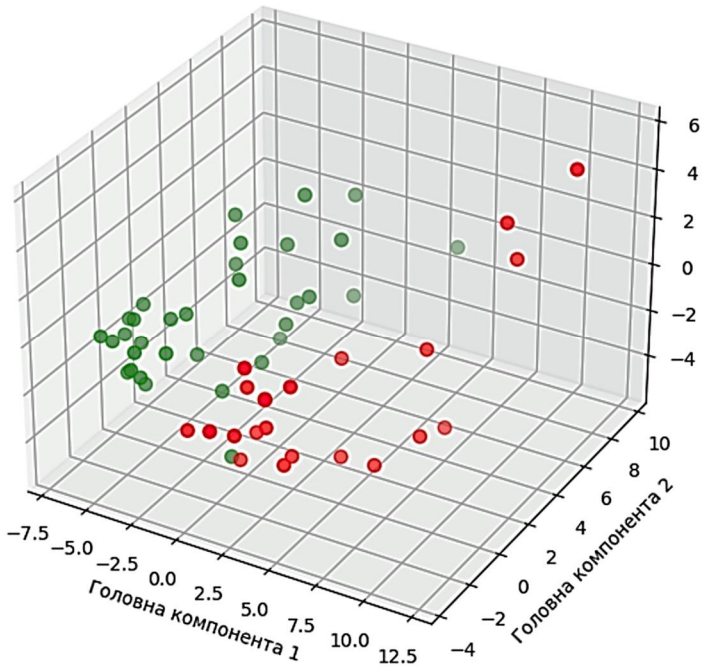

Рис. 3. Графік PCA для розмічених даних

На графіку видно, що розташування точок виявляє певні кластери, що сприяє візуальному аналізу та оцінці групування даних за класами. Це надає можливість краще зрозуміти, які ознаки можуть бути корисними для подальшої класифікації даних.

Для покращення результатів точності моделі для нерозмічених даних необхідно шукати та застосовувати альтернативні методи аналізу й опрацювання таких даних або поліпшувати алгоритми кластеризації.

Висновки

У статті розглянуто наявні інформаційні системи напівконтрольованого навчання, а також проаналізовано їхню структуру та результати роботи. На основі цього аналізу розроблено власну інформаційну систему напівконтрольованого навчання, яка здатна ефективно працювати і з розміченими, і з нерозміченими даними. Нова система демонструє здатність аналізувати та класифікувати дані, що є важливим для різних медичних застосувань.

Метрики, отримані зі звіту, вказують на те, що система потребує подальшого вдосконалення для досягнення більшої точності. Незважаючи на це, поточні результати вже дають змогу використовувати цю систему в практичній діяльності лікарів-діагностів. Інформаційна система може значно допомогти у прийнятті обґрунтованих рішень, забезпечуючи швидший доступ до даних і підвищуючи точність діагностики. Завдяки автоматизації процесів система допомагає уникати людських помилок, що сприяє покращенню безпеки пацієнтів.

Отже, розроблена інформаційна система напівконтрольованого навчання є перспективним інструментом для медичної діагностики. Вона може значно підвищити ефективність роботи медичних фахівців, забезпечуючи більш точні та швидкі діагностичні рішення. Подальше вдосконалення цієї системи може призвести до ще кращих результатів, а це сприятиме покращенню якості медичних послуг і безпеки пацієнтів.

СПИСОК ЛІТЕРАТУРИ

1. Xu, L., Abidi, S. R. (2019) Intelligent health data analytics: A convergence of artificial intel- ligence and big data. *Healthcare Management Forum, 32*(4), 178–182. https://doi.org/doi.org/10.1177/0840470419846134.

2. Wang, X., Calvanese, D. (2021, July). Editorial for Special Issue of Journal of Big Data Research on “Big Data Meets Knowledge Graphs”. *Big Data Research, 25,* 202–215. https://doi.org/10.1016/j.bdr.2021.100215.

3. Philip Chen, C., Han, R. (2022, June). Graph-based sparse bayesian broad learning system for semi-supervised learning. *Information Sciences*, *597,* 193–210. https://doi.org/10.1016/j.ins.2022.03.037.

4. Tran, Q. T., Alom, M. Z., & Orr, B. A. (2022, June 8). Comprehensive study of semi-supervised learning for DNA methylation-based supervised classification of central nervous system tumors. *BMC Bioinformatics*, *23*(1), 313-319. https://doi.org/10.1186/s12859-022-04764-1.

5. Yadav, S. K., S., V. (2019, April 30). EEG Classification using Semi Supervised Learning. *International Journal of Trend in Scientific Research and Development*, *3*(3), 1441–1445. https://doi.org/10.31142/ijtsrd23355.

6. Nasrabadi, N. M. (2007). Pattern Recognition and Machine Learning. *Journal of Electronic Imaging, 16*(4), 172–178. https://doi.org/10.1117/1.2819119

7. Nazirova, T. O., Kostenko, O. B. (2018, October 25). Нейромережева інформаційна технологія опрацювання медичних даних. *Scientific Bulletin of UNFU, 28*(8), 141–145. https://doi.org/10.15421/40280828.

8. Lyrchykov, V. O., Baybuz, O. H. (2022, December 25). Технологія видобутку даних про ризики захворювання на основі аналізу електронних медичних карток. *Actual Problems of Automation and Information Technology, 26*(1), 118–129. https://doi.org/10.15421/432208.

9. Nazirova, T. O., Kostenko, O. B. (2018, October 25). Нейромережева інформаційна технологія опрацювання медичних даних. *Scientific Bulletin of UNFU, 28*(8), 141–145. https://doi.org/10.15421/40280828.

10. Ключко, О. (2020, July 13). Електронні інформаційні системи в медицині та в біології: загальний аналіз. *Medical Informatics and Engineering, 2,* 111–123. https://doi.org/10.11603/mie.1996-1960.2020.2.11183.

11. Гумен, О., Рачек, К. (2023, December 26). Нейронні мережі та машинне навчання у обробці даних для прогнозування космічної погоди. *Applied Questions of Mathematical Modeling, 6*(2), 19–23. https://doi.org/10.32782/mathematical-modelling/2023-6-2-2.

REFERENCES

1. Xu, L., Abidi, S. R. (2019) Intelligent health data analytics: A convergence of artificial intel- ligence and big data. *Healthcare Management Forum, 32*(4), 178–182. https://doi.org/doi.org/10.1177/0840470419846134.

2. Wang, X., Calvanese, D. (2021, July). Editorial for Special Issue of Journal of Big Data Research on “Big Data Meets Knowledge Graphs”. *Big Data Research, 25,* 202–215. https://doi.org/10.1016/j.bdr.2021.100215.

3. Philip Chen, C., Han, R. (2022, June). Graph-based sparse bayesian broad learning system for semi-supervised learning. *Information Sciences*, *597*, 193–210. https://doi.org/10.1016/j.ins.2022.03.037.

4. Tran, Q. T., Alom, M. Z., Orr, B. A. (2022, June 8). Comprehensive study of semi-supervised learning for DNA methylation-based supervised classification of central nervous system tumors. *BMC Bioinformatics*, *23*(1), 313-319. https://doi.org/10.1186/s12859-022-04764-1.

5. Yadav, S. K., S., V. (2019, April 30). EEG Classification using Semi Supervised Learning. *International Journal of Trend in Scientific Research and Development*, *3*(3), 1441–1445. https://doi.org/10.31142/ijtsrd23355.

6. Nasrabadi, N. M. (2007). Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*, *16*(4), 172–178. https://doi.org/10.1117/1.2819119

7. Nazirova, T. O., Kostenko, O. B. (2018, October 25). Neural network information technology for processing medical data. *Scientific Bulletin of UNFU, 28*(8), 141–145. https://doi.org/10.15421/40280828.

8. Lyrchykov, V. O., Baybuz, O. H. (2022, December 25). Technology of extracting data on disease risks based on the analysis of electronic medical records. *Actual Problems of Automation and Information Technology*, *26*(1), 118–129. https://doi.org/10.15421/432208.

9. Nazirova, T. O., Kostenko, O. B. (2018, October 25). Neural network information technology for processing medical data. *Scientific Bulletin of UNFU, 28*(8), 141–145. https://doi.org/10.15421/40280828.

10. Kliuchko, O. (2020, July 13). Electronic information systems in medicine and biology: a general analysis. *Medical Informatics and Engineering, 2,* 111–123. https://doi.org/10.11603/mie.1996-1960.2020.2.11183.

11. Humen, O., Rachek, K. (2023, December 26). Neural networks and machine learning in data processing for space weather forecasting. *Applied Questions of Mathematical Modeling*, *6*(2), 19–23. https://doi.org/10.32782/mathematical-modelling/2023-6-2-2

INFORMATION SYSTEM OF SEMI-SUPERVISED LEARNING FOR ANALYSIS OF DATA SAMPLES WITH HIGH DIMENSIONS

Nelia Miroshnychenko[1], Iryna Perova,[2] Oleg Datsok[3]

[1-3] Kharkiv National University of Radio Electronics,
Department of Systems Engineering, Kharkiv, Ukraine
[1] E-mail: nelia.miroshnychenko@nure.ua, ORCID: 0000-0002-3846-1668
[2] E-mail: iryna.perova@nure.ua, ORCID: 0000-0003-2089-5609
[3] E-mail: oleh.datsok@nure.ua, ORCID: 0000-0003-4489-3819

The study of large datasets to uncover hidden patterns and trends has become increasingly important and valuable in recent years. These large datasets are characterized by wide availability, structural complexity, and significant volume of information.

This article proposes a detailed description of a semi-supervised learning information system for analyzing high-dimensional data samples. The system is designed to process large datasets using semi-supervised learning methods for effective analysis and classification. Existing information systems capable of working with high-dimensional data samples, as well as methods for efficient analysis and classification of these data samples, were analyzed for this purpose.

The article provides a detailed description of the system architecture, including data processing methods, feature selection, preprocessing modules, and training optimization methods.

Keywords: information system of semi-supervised learning, feature extraction, large datasets, group search of features, fuzzy adaptive Kohonen network.