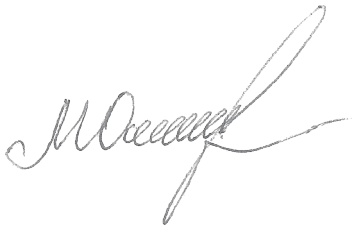
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ ПОЛІТЕХНІЧНИЙ УНІВЕРСИТЕТ

УДК 004.89:316



Манікаєва Ольга Сергіївна

МЕТОДИ КЛАСИФІКАЦІЇ ГЕТЕРОГЕННИХ ДАНИХ В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ МЕДИКО-СОЦІОЛОГІЧНОГО МОНІТОРИНГУ

05.13.23 – Системи та засоби штучного інтелекту

Автореферат дисертації
на здобуття наукового ступеня
кандидата технічних наук

Одеса – 2020

Дисертацією є рукопис.

Робота виконана в Одеському національному політехнічному університеті Міністерства освіти і науки України.

Науковий керівник — доктор технічних наук, професор **Арсірій Олена Олександрівна**, Одеський національний політехнічний університет, завідувач кафедри інформаційних систем

Офіційні опоненти: — доктор технічних наук, професор **Литвиненко Володимир Іванович**, Херсонський національний технічний університет, завідувач кафедри інформаційних технологій і комп'ютерних наук

доктор технічних наук, професор **Субботін Сергій Олександрович** Національний університет «Запорізька політехніка», завідувач кафедри програмних засобів

Захист відбудеться 26 січня 2021 р. о 13:30 на засіданні спеціалізованої вченої ради К 41.052.08 в Одеському національному політехнічному університеті за адресою: 65044, м. Одеса, пр. Шевченка, 1, ауд. 204.

З дисертацією можна ознайомитися у бібліотеці Одеського національного політехнічного університету за адресою: 65044, м. Одеса, пр. Шевченка, 1.

Автореферат розісланий 24 грудня 2020 р.

Вчений секретар
спеціалізованої вченої ради — Р. О. Шапорін

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Під час ускладнення світової епідеміологічної ситуації особлива увага приділяється спостереженням за станом умов життя та праці суспільства на основі розробки інтелектуальних систем (ІС) медико-соціологічного моніторингу (МСМ) нового покоління. Такі ІС МСМ в *оперативному режимі* дозволяють не тільки збирати та накопичувати *гетерогенні* дані про стан громадського здоров'я або мікроклімат соціального середовища досліджуваної цільової аудиторії, а й приймати класифікаційні рішення про образ життя або умови праці кожного із респондентів. Аналіз існуючих ІС МСМ, таких як «МедЕксперт», «Ексімед», «Meditex», «Myrian», «Centrak» та інш., показує, що для класифікаційних рішень в таких системах все частіше використовують методи навчання глибинних нейронних мереж зі вчителем. Але практичне використання таких методів потребує тривалої попередньої обробки *слабко структурованих* гетерогенних даних МСМ, які не тільки збираються з різних джерел, а і інтерпретуються за допомогою різних, не завжди пов'язаних, якісних або кількісних шкал та часто суперечать одне одному. Цей факт значно знижує необхідну в умовах моніторингу швидкість або достовірність прийняття класифікаційних рішень. Крім того, в ситуації коли слабко структуровані гетерогенні дані МСМ періодично оновлюються та поповнюються, неможливо отримати достатню кількість класифікаційних рішень у вигляді попередньо розмічених даних. Тому гетерогенні дані МСМ можна віднести також до *слабко маркованих.*

Слід відзначити, що для класифікації таких гетерогенних даних використовують *навчання нейронних мереж з частковим залученням вчителя,* яке може бути реалізовано за допомогою самонавчання нейронної мережі Кохонена на немаркованих даних з метою подальшого їх додавання до навчальної вибірки. В цьому випадку швидкість класифікації даних МСМ залежить від швидкості самонавчання мережі Кохонена, після проведення якого необхідно вирішувати ще й *завдання узгодження наявних попередньо маркованих даних із отриманими маркерами кластерів*. Крім того коригування вагових коефіцієнтів нейронів за алгоритмом WTA з градієнтною оптимізацією в процесі навчання мережі Кохонена на навчальній вибірці, яка складається із суперечливих слабко маркованих даних, потребує значних затрат часу, що є не прийнятним. Тому використання для коригування вагових коефіцієнтів нейронних мереж *генетичних алгоритмів* з притаманною їм можливістю випадкової мутації дозволяє знизити негативний вплив помилкових рішень на загальний рівень швидкості навчання та достовірності класифікації.

Таким чином, є *актуальним* вирішення науково-практичної задачі розробки моделей представлення слабко структурованих та слабко маркованих гетерогенних даних МСМ та методів їх класифікації на основі навчання мережі Кохонена з використанням генетичних алгоритмів для створення ІС МСМ. Практичне використання розробленої ІС МСМ дозволить в оперативному режимі приймати достовірні класифікаційні рішення про образ життя або умови праці кожного із респондентів досліджуваної цільової аудиторії.

Зв’язок роботи з науковими програмами, планами і темами. В основу дисертації покладено результати досліджень і розробок, одержаних під час

виконання науково-дослідних робіт на кафедрах інформаційних систем в менеджменті та інформаційних систем Одеського національного політехнічного університету: № 700-145 «Моделі, методи та інструментальні засоби підтримки прийняття рішень з підвищення ефективності гідроаеродинамічних процесів в діючому енергетичному обладнанні» (№ ДР 0115U000413); № 448-68 «Дослідження інформаційних сховищ як моделей предметних областей в системах підтримки прийняття рішень» (№ ДР 0104U002401); у проекті Європейського Союзу Erasmus+ «Methods and models of neural networks for pattern recognition to decision support system of monitoring the functional state of a complex system», грантова угода № 2015-1-ES01-K107-015381.

Мета і задачі дослідження. Метою дисертаційної роботи є зниження часу на прийняття класифікаційних рішень без втрати рівня їх достовірності в інтелектуальних системах медико-соціологічного моніторингу за рахунок розроблення моделей представлення слабко структурованих та слабко маркованих гетерогенних даних та методів їх класифікації на основі мережі Кохонена.

Для досягнення поставленої мети в роботі вирішено наступні задачі:

– проаналізовано методи класифікації гетерогенних даних в існуючих ІС МСМ та обрано напрямки дослідження;

– розроблено моделі представлення слабко структурованих та слабко маркованих гетерогенних даних МСМ;

– розроблено методи класифікації слабко структурованих та слабко маркованих гетерогенних даних МСМ;

– розроблено ІС МСМ та проведено її апробацію з позитивним результатом.

Об’єкт дослідження – процес класифікації гетерогенних даних в інтелектуальних системах медико-соціологічного моніторингу.

Предмет дослідження – моделі представлення та методи класифікації слабко структурованих та слабко маркованих гетерогенних даних в інтелектуальних системах медико-соціологічного моніторингу.

Методи дослідження. При розв’язанні поставлених у дисертаційній роботі задач використано методи статистичного аналізу для удосконалення моделей представлення слабко структурованих та слабко маркованих гетерогенних даних МСМ в просторах властивостей та ознак, методи дискримінантного, кластерного аналізу та еволюційних обчислювань для розробки методу навчання мережі Кохонена для класифікації даних МСМ, принципи об’єктно-орієнтованого програмування при розробці ІС МСМ, методи експериментальних досліджень для її апробації.

Наукова новизна одержаних результатів полягає у вирішенні науково-практичної задачі розробки моделей представлення слабко структурованих та слабко маркованих гетерогенних даних медико-соціологічного моніторингу та методів їх класифікації на основі навчання мережі Кохонена з використанням генетичних алгоритмів.

1. Удосконалено моделі представлення слабко структурованих та слабко маркованих гетерогенних даних МСМ в просторах властивостей та ознак з урахуванням значень, типів, форматів, джерел, оцінок якості та процедури агрегування/трансформації властивостей деталізованих даних, що дозволило

створювати навчальну та тестову вибірку для подальшої класифікації з частковим залученням вчителя.

2. Вперше розроблено метод узгодження маркерів класів та кластерів при навчанні мережі Кохонена з частковим залученням вчителя, який заснований на побудові двовимірної гістограми попарних збігів значень маркерів класів і кластерів з її подальшим коригуванням за рядками та за стовпцями у відповідності до розробленого правила, що дозволило отримати додаткові маркери класів в немаркованій частині навчальної вибірки.

3. Отримав подальший розвиток метод евристичного коригування вагових коефіцієнтів в процесі навчання нейронної мережі Кохонена за рахунок використання методу узгодження маркерів класів та кластерів для оцінки пристосованості хромосоми в популяції генетичного алгоритму для кожного прикладу із навчальної вибірки, що дозволило зменшити час навчання мережі Кохонена без втрати рівня достовірності класифікації.

4. Удосконалено метод класифікації слабко структурованих та слабко маркованих гетерогенних даних МСМ за рахунок евристичного коригування вагових коефіцієнтів в процесі навчання мережі Кохонена, що дозволило знизити час класифікації в ІС МСМ без зниження рівня її достовірності.

Практичне значення одержаних результатів полягає у створенні на основі запропонованих рішень ІС МСМ, яка дозволить накопичувати та зберігати гетерогенні дані МСМ та в оперативному режимі приймати достовірні класифікаційні рішення про образ життя або умови праці кожного із респондентів досліджуваної цільової аудиторії при проведенні МСМ. ІС МСМ складається із двох підсистем: *підготовки та зберігання вхідних даних*, яку побудовано за принципом ETL-систем з використанням розроблених моделей деталізованих та агрегованих даних та *прийняття класифікаційних рішень* на основі мережі Кохонена, для навчання якої застосовано удосконалені методи.

Розроблену ІС МСМ апробовано на елеваторному комплексі ТОВ «Сервіс-Грейн» (акт впровадження від 02.08.2020) при впровадженні проекту з охорони праці на підприємстві, що дозволило покращити умови праці в середньому на 15%. Досягнуто значне скорочення часу на обстеження умов праці за рахунок зменшення вдвічі часу класифікації рівнів аерозольного, електромагнітного, акустичного, хімічного і біологічного впливу, іонізуючого випромінювання, мікроклімату, освітленості і вібрації на підприємстві.

Використання розробленої ІС МСМ під час проведення комплексної експертизи медико-соціального становища студентської аудиторії на кафедрі медичної інженерії Львівського національного медичного університету ім. Данила Галицького (акт впровадження від 02.11.2020), дозволило автоматизувати процес обробки вхідних даних та на 50% знизити час прийняття класифікаційних рішень.

Основні положення, висновки та рекомендації, що викладені в дисертаційній роботі використовуються в навчальному процесі Одеського національного політехнічного університету в дисциплінах «Методи та системи штучного інтелекту», «Інтелектуальний аналіз даних» та «Прикладні аспекти глибинного навчання», які викладаються на кафедрі інформаційних систем для студентів спеціальності 122 – «Комп’ютерні науки».

Особистий внесок здобувача. Особистий внесок здобувача полягає в удосконаленні моделей представлення слабко структурованих та слабко маркованих гетерогенних даних МСМ [6,7]; в розробці методу узгодження маркерів класів та кластерів при навчанні мережі Кохонена з частковим залученням вчителя [1,2,3,8,9,10,11,12]; в подальшому розвитку методу евристичного коригування вагових коефіцієнтів в процесі навчання нейронної мережі Кохонена за рахунок використання методу узгодження маркерів класів та кластерів для оцінки пристосованості хромосоми в популяції генетичного алгоритму для кожного прикладу із навчальної вибірки [5,13]; в удосконаленні методу класифікації слабко структурованих та слабко маркованих гетерогенних даних МСМ [4,7].

Апробація результатів дисертації. Основні положення дисертаційної роботи доповідалися та обговорювалися на наукових семінарах кафедри інформаційних систем ОНПУ та наукових конференціях: V Міжнародна науково-практична конференція магістрантів, аспірантів та науковців «Управління проектами: інновації, нелінійність, синергетика» (м. Одеса, 2014 р.); III та V Українсько-німецька конференція «Інформатика. Культура. Техніка» (м. Одеса, 2015, 2017 рр.); IV Міжнародна науково-практична конференція «Інформаційні системи та технології» (м. Харків, 2015 р.); XVII Міжнародна науково-практична конференція «Сучасні інформаційні та електронні технології» (м. Одеса, 2016 р.); VI Міжнародна науково-практична конференція «Інформаційні управляючі системи та технології» (м. Одеса, 2017 р.).

Публікації. За темою дисертації опубліковано 13 наукових праць, серед них 5 статей у фахових наукових виданнях із переліку МОН України, що включені до міжнародних наукометричних баз, з них 2 статті у виданнях, що індексуються у наукометричній базі даних SCOPUS та 7 праць або тез доповідей у міжнародних та всеукраїнських науково-практичних конференціях.

Структура дисертаційної роботи. Дисертація складається зі вступу, 4 розділів, висновків, списку використаних джерел із 108 назв і 3 додатків. Загальний обсяг дисертації становить 125 сторінок. Робота містить 52 рисунка і 9 таблиць.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету, задачі, об’єкт, предмет та методи дослідження, визначено наукову новизну та практичне значення одержаних результатів, наведено відомості про публікації та апробацію результатів дисертаційної роботи.

У **першому розділі** проведено аналіз спеціалізованих інтелектуальних ІС для обробки гетерогенних даних МСМ. Показано, що процес прийняття рішення в ІС МСМ складається із чотирьох етапів: *формування гіпотез дослідження* про склад цільової аудиторії (соціального середовища), *підготовки і зберігання* даних МСМ, *прийняття класифікаційних рішень* для підтвердження або спростування гіпотез дослідження, а також *інтерпретації/оцінки* отриманих результатів класифікації (рис. 1). Встановлено, що інтелектуальна діяльність експерта-соціолога щодо формування гіпотез дослідження та прийняття класифікаційних рішень про стан громадського здоров’я цільової аудиторії на основі аналізу кількісних метрик або якісних оцінок соціально-демографічних, медичних, психофізичних, поведінкових,

географічних або будь-яких інших характеристик пов’язана з узагальненням даних, що складаються із самооцінок населенням свого здоров'я, якості життя, задоволеності медичними та соціальними послугами. Такі дані МСМ є гетерогенними тому, що періодично поповнюються із заповнених форм опитувань, анкет та інтерв'ю або електронних поштових повідомлень, візиток, pdf- і txt-файлів, миттєвих повідомлень в різних месенджерах (WhatsApp, Viber, Facebook Messenger, Skype, ICQ, Google Hangouts та ін.), документів, веб-сторінок, рахунків, аудіо / відео, чеків і контрактів, картинок та ін. Крім того, вхідні гетерогенні дані МСМ є *слабко структурованими* (*СС*), тому що збираються з різних джерел, інтерпретуються за допомогою різних і не завжди пов'язаних шкал і часто суперечать одне одному. Відомо, що аналіз великих масивів слабко структурованих гетерогенних даних потребує *тривалої в часі ручної праці* кваліфікованих експертів соціологів, тому для вхідних даних МСМ, які періодично поповнюються, такий процес прийняття класифікаційних рішень є в деяких випадках зовсім неможливим.

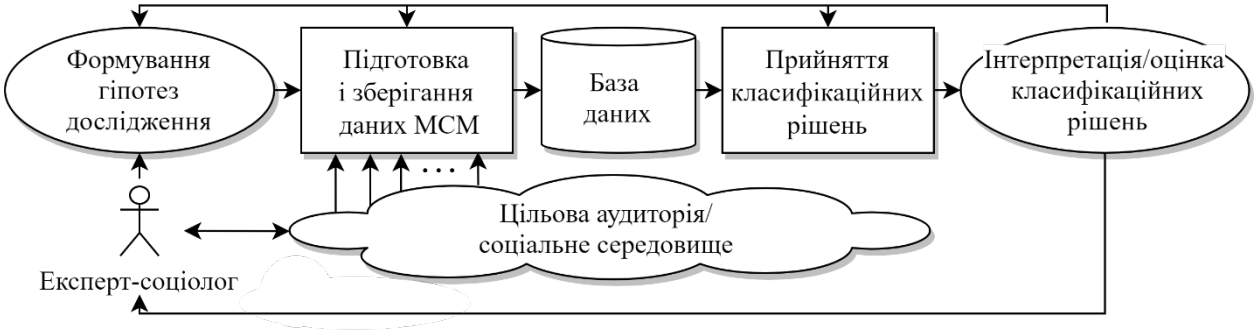


Рисунок 1 – Загальна структура типової ІС МСМ

З іншого боку показано, що створення ІС МСМ на основі машинного навчання зі вчителем (*supervised learning*) дозволяє автоматизувати процес отримання класифікаційних рішень про цільову аудиторію тільки при наявності попередньо маркованих (*labeled*) емпіричних кількісних або категоріальних даних. Однак, попередньо встановлено, що отримання достатньої кількості таких маркованих даних (навчальної вибірки) в випадку класифікації слабко структурованих гетерогенних даних МСМ є неможливим. Тому на практиці гетерогенні дані МСМ є *слабко маркованими* (*СМ*) і для отримання класифікаційних рішень при обробці таких даних використовують методи машинного навчання з частковим залученням вчителя (*semi-supervised learning SSL*), які крім маркованих даних для навчання використовують також зазвичай порівняно більшу кількість немаркованих (*unlabeled*) даних. Дослідження показують, що одним із поширених способів реалізації методів SSL є використання самонавчання (*self-learning*) нейронної мережі Кохонена з конкуруючим шаром на немаркованих даних з метою подальшого їх додавання до навчальної вибірки. Але в цьому випадку швидкість класифікації даних МСМ залежить від швидкості самонавчання мережі Кохонена, після проведення якого необхідно вирішувати ще й завдання узгодження наявних попередньо маркованих даних із отриманими маркерами кластерів. Крім того ситуацію з необхідною швидкістю класифікації даних МСМ ускладнює той факт, що коригування вагових коефіцієнтів нейронів за алгоритмом WTA з градієнтною оптимізацією в процесі навчання мережі Кохонена на навчальній вибірці, яка складається із суперечливих *СМ* даних, потребує значних затрат часу.

З іншого боку відомо, що у випадку коли досліджуються великі області допустимих значень параметрів нейронних мереж, а поверхня помилки навчання має складний рельєф, використання для навчання нейронних мереж *генетичних алгоритмів* (*ГА*) з притаманною їм можливістю випадкової мутації, дозволяє знизити негативний вплив помилкових рішень на загальний рівень достовірності класифікації. З урахуванням необхідності вирішення задачі класифікації слабко структурованих та слабко маркованих (*CC&CM*) гетерогенних даних в ІС МСМ сформульовано мету та задачі дослідження.

У **другому розділі** удосконалено моделі представлення *CC&CM* гетерогенних даних МСМ для формалізації побудови навчальної та тестової вибірок та метод узгодження маркерів класів та кластерів при навчанні мережі Кохонена з частковим залученням вчителя для отримання додаткових маркерів класів в немаркованій частині навчальної вибірки.

Удосконалення моделей представлення *CC&CM* гетерогенних даних МСМ. Задачу класифікації респондентів цільової аудиторії в ІС МСМ формалізуємо наступним чином. Нехай цільова аудиторія (*Target Audience,* TA) складається з множини респондентів (*Respondents,* R_i):

$$TA = \{R_1, R_2, \dots, R_i, \dots, R_n\}, i = \overline{1,n}, \quad (1)$$

де R_i – респондент, що досліджується, n – кількість респондентів.

Згідно зі структурою типової ІС МСМ (див. рис. 1) експерт-соціолог висуває гіпотезу про те, що кожен респондент R_i цільової аудиторії TA належить до класу t^k із множини значень класів:

$$T = \{t^1, \dots, t^k, \dots, t^p\}, k = \overline{1,p}, \quad (2)$$

де p – потужність множини класів $|T|$. При цьому $p \ll n$.

З іншого боку, для прийняття класифікаційних рішень, щодо респондентів в ІС (див. рис. 1) використовують набори *CC* гетерогенних даних МСМ, вид, тип та формат яких залежить від способу отримання вхідної інформації. Дані про респондента R_i, що вилучені із різних джерел інформації є *деталізованими даними*. Тоді, враховуючи представлення (1), запропоновано деталізовані дані D_{R_i} в m-вимірному просторі властивостей задавати за допомогою наступної моделі:

$$D_{R_i} = \langle \{V_\varphi\}_j, T_j, F_j, S_j, Q_j, Mt_j \rangle, j = \overline{1,q}, \quad (3)$$

де $\{V_\varphi\}_j, T_j, F_j, S_j, Q_j, Mt_j$ – множина дискретних (якісних або кількісних) значень, тип, формат, джерело, оцінка якості, процедура агрегування/трансформації даних j-тої властивості; q – кількість властивостей респондента R_i . При цьому $q \ll n$.

Для зниження часу на прийняття класифікаційного рішення t^k без втрати рівня достовірності, множина дискретних значень $\{V_\varphi\}_j$ деталізованих даних D_{R_i} кожного типу з урахуванням формату представлення за допомогою процедури Mt_j агрегується та трансформується в простір ознак:

$$x_j = Mt_j(\{V_\varphi\}_j), \quad (4)$$

де процедура Mt_j в залежності від типу, формату, джерела та оцінки якості отриманих дискретних значень $\{V_\varphi\}_j$ обирається із множини можливих функцій квантування (Quant), сортування (Sort), злиття (Merg), групування (Group), налаштування набору даних (DSett), табличної підстановки значень (TSubst),

обчислювання значень (CalcVal), кодування даних (DEncod), нормалізації (масштабування) (Norm(Scal)) та інш., таким чином:

$$Mt_j \in \{\text{Quant, Sort, Merg, Group, DSett, TSubst, CalcVal, DEncod, Norm(Scal)}\}. \quad (5)$$

Тоді модель агрегованих даних A_{R_i} респондента R_i (1) в q-вимірному просторі ознак з врахуванням представлень (2) – (4) задається за допомогою кортежу:

$$A_{R_i} = \langle x_1, \dots, x_j, \dots, x_q, t_i^k \rangle,\ j = \overline{1,q},\ k = \overline{1,p}, \quad (6)$$

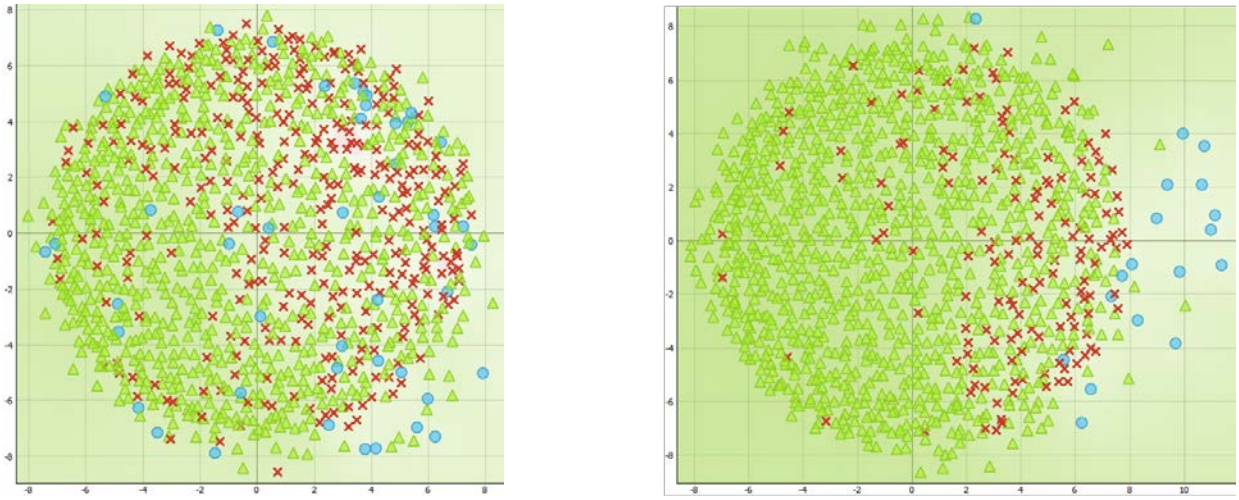
де q – розмірність простору ознак, а t_i^k – маркер класу k із множини T (2), до якого належить i-респондент. Маркер класу t_i^k визначається в ручному або автоматизованому режимі трьома шляхами:

1) за власним рішенням респондента R_i тільки в випадках проведення опитування, анкетування або інтерв'ю – ручний режим;

2) в результаті аналізу значень множини агрегованих даних $A_{R_i} = \{x_j\}$, який проводить експерт-соціолог на підставі власного досвіду і знань – ручний режим;

3) в результаті побудови класифікаційного правила $a: X \to T$ за допомогою машинного навчання зі вчителем, в якості якого використовується навчальна вибірка $A_R = \{A_{R_i}\}$ – автоматизований режим.

На рисунку 2 показаний точковий графік у вигляді карти подібності респондентів МСМ «Україна – стиль життя», який побудовано за трьохетапним сценарієм. На першому етапі за 114 властивостями для 1143 респондентів вилучено вхідні деталізовані дані D_R, що характеризують стиль життя за *трьома* класами. На другому етапі дані D_R згідно (3) – (4) агреговано та трансформовано у простір 23 ознак, при цьому отримано $A_{R_i} = \{x_j\}$ $j = \overline{1,23}; k = \overline{1,3}$. На третьому етапі A_R, за допомогою багатовимірного шкалювання (масштабування), трансформовано у двовимірний простір так званих *Multidimensional scaling – MDS* проекцій, які відображено з урахуванням маркерів класів t_i^k, що отримані в ручному режимі за рішенням респондента (рис. 2,*а*), або експерта (рис. 2,*б*).



– «погане», – «середнє», – «добре»

а *б*

Рисунок 2 – Карта подібності респондентів МСМ за стилем життя для маркерів класів q=3 *а*) за власним рішенням респондента; *б*) за рішенням експерта-соціолога

Таким чином показано, що на створення навчальної вибірки A_{R_i}, впливають суб'єктивні рішення респондентів, які залежать від попередньо висунутої експертом-соціологом гіпотези про потужність множини класів $|T|$, та/або рішення експерта-соціолога про відношення респондента до певного класу. Встановлено, що процес прийняття таких рішень є неоднозначним і довготривалим, вимагає інтелектуальної праці експерта соціолога та залежить від його кваліфікації. Що в свою чергу свідчить, що в типовій ІС МСМ (див. рис. 1) для *CC* гетерогенних даних МСМ, які ще і постійно оновлюються, є практично неможливим отримати достатню кількість попередньо маркованих даних для прийняття в автоматизованому режимі достовірних класифікаційних рішень.

Тому відповідно до (1) і (6) в матричному вигляді навчальна вибірка (*training set*) даних МСМ є набором агрегованих даних респондентів R_i цільової аудиторії TA та складається із маркованих (*label*) та немаркованих (*unlabeled*) даних:

$$A_R^{train} = A_R^l \cup A_R^u, \tag{7}$$

де $$A_R^l = \langle X, T \rangle = \langle \begin{bmatrix} x_{11} & \dots & x_{1q} \\ \dots & \dots & \dots \\ x_{l1} & \dots & x_{lq} \end{bmatrix}, \begin{bmatrix} t_1^1 \\ \dots \\ t_l^p \end{bmatrix} \rangle, A_R^u = \langle X \rangle = \langle \begin{bmatrix} x_{l+11} & \dots & x_{l+1q} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nq} \end{bmatrix} \rangle, \tag{8}$$

де матриця X – множина ознак респондентів R_i в q-вимірному просторі ознак, а вектор T – кінцева множина номерів (імен, міток, маркерів) класів (2).

Тестова вибірка (*test set*) має вигляд:

$$A_R^{test} = \langle \begin{bmatrix} x_{n+11} & \dots & x_{n+1q} \\ \dots & \dots & \dots \\ x_{g1} & \dots & x_{gq} \end{bmatrix} \rangle. \tag{9}$$

Тоді задачу класифікації даних МСМ в ІС, в термінах машинного навчання з частковим залученням вчителя, запропоновано формалізувати наступним чином: існує невідома *цільова залежність* – $a: X \to T$, значення якої відомі тільки для кінцевого числа респондентів R_i із маркованої частини навчальної вибірки A_R^l. З урахуванням даних із немаркованої частини навчальної вибірки A_R^u потрібно побудувати алгоритм $a^*: X \to T$, здатний класифікувати будь-якого респондента $R_i \in A_R^{test}$ із заданою достовірністю. Для класифікації даних МСМ з частковим залученням вчителя, запропоновано за допомогою нейронної мережі Кохонена спочатку провести кластерний аналіз, в рамках якого вирішується задача розбиття навчальної вибірки A_R^{train} даних МСМ на непересічні кластери:

$$Z = \{z^1, \dots, z^m, \dots, z^p\}, \; m = \overline{1, p}, \tag{10}$$

де z^m – маркер кластера, що містить вектори, що схожі за ознаками з вектором A_{R_i} (6) з навчальної вибірки A_R^{train}, $i = \overline{1, n}$, де n – кількість векторів у вибірці:

$$z^m = \left\{ A_{R_i}, A_{R_j} \middle| \begin{matrix} A_{R_i} \in A_R^{train}, A_{R_j} \in A_R^{train}, \\ d(A_{R_i}, A_j) < \sigma \end{matrix} \right\}, \tag{11}$$

де $d(A_{R_i}, A_{R_j})$ – міра близькості між респондентами, що називається відстанню; σ – величина, що визначає поріг близькості для включення респондентів в один кластер.

Отриманий набір кластеризованих даних $A_R^{train^*}$ має вигляд:

$$A_R^{train^*} = \langle X, Z \rangle \langle \begin{bmatrix} x_{11} & \dots & x_{1q} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nq} \end{bmatrix}, \begin{bmatrix} z_1^1 \\ \dots \\ z_n^p \end{bmatrix} \rangle. \quad (12)$$

Розробка методу узгодження маркерів класів та кластерів при навчанні мережі Кохонена з частковим залученням вчителя. Запропоновано автоматизувати процес навчання мережі Кохонена з використанням *СМ* гетерогенних даних МСМ A_R^{train} (7) за допомогою методу, що включає п’ять етапів (рис. 3).

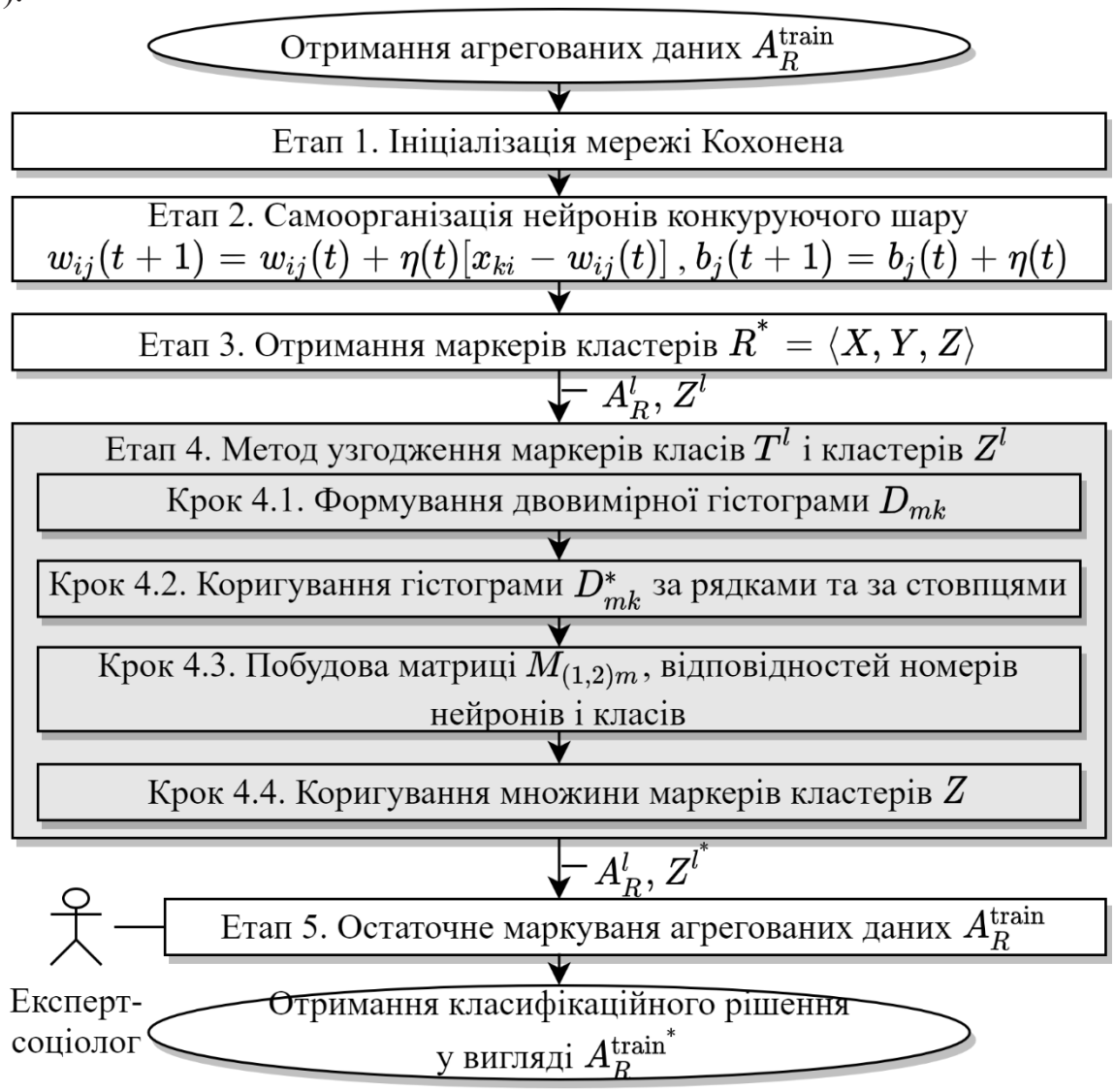


Рисунок 3 – Етапи методу навчання мережі Кохонена з частковим залученням вчителя

Етап 1. У відповідності до розмірів простору ознак q та потужності множини класів p створюється конкуруючий шар мережі Кохонена та з врахуванням значень A_R^{train} (7) проводиться ініціалізація матриці вагових коефіцієнтів $\{w_{ij}\}$ нейронів конкуруючого шару та їх зміщень $\{b_j\}$.

Етап 2. Процедура самоорганізації конкуруючого шару реалізована на основі класичного ітеративного алгоритму WTA. На вхід конкуруючого шару послідовно подаються значення векторів $\{x_{ki}\}$ з навчальної вибірки A_R^{train}. Визначення нейрона переможця відбувається за мінімізацією відстаней: $d(x_{ki}, w_{ij}) \rightarrow \min d(x_{ki}, w_{ij})$. Як

міра відстані використовується евклідова відстань $d(x_{ki}, w_{ij}) = \sqrt{\sum_{i=1}^{q}(x_{ki} - w_{ij})^2}$ між елементами вхідних векторів x_{ki} і ваговими коефіцієнтами w_{ij}.

Етап 3. Для отримання маркерів кластерів на вхід мережі Кохонена отриманими на етапі 2 матрицею W та вектором B з навчальної вибірки A_R^{train} послідовно подаються значення векторів $\{x_{ki}\}$, $i = \overline{1,q}$, q – розмірність простору ознак. Відклик мережі формується у вигляді бінарного вектору $\{y_{kj}\}$, $j = \overline{1,p}$, де p – число нейронів конкуруючого шару – для k-го вектору. При цьому y_{kj}= 1, якщо j є номером нейрона переможця і y_{kj}= 0 – у всіх інших випадках.

Результатом етапу є вихідний вектор $\{z_{kj}\}$, значенню k-того елемента якого присвоюється порядковий номер нейрона-переможця j-того бінарного вектору $\{y_{kj}\}$.

Таким чином отримаємо попереднє рішення в наступному вигляді:

$$R^* = \langle X, Y, Z \rangle \left\langle \begin{bmatrix} x_{11} & \cdots & x_{1q} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nq} \end{bmatrix}, \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \cdots & \cdots & \cdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix}, \begin{bmatrix} z_1^1 \\ \cdots \\ z_n^p \end{bmatrix} \right\rangle, \quad (13)$$

де X – множина ознак респондентів R_i, Y – бінарна матриця відкликів конкуруючого шару, Z – множина маркерів кластерів.

Етап 4. Для *узгодження існуючих маркерів* класів множини T (2) до якої належить респондент R_i з маркованих даних A_R^l (8) із *значеннями маркерів кластерів* множини Z (13) розроблено відповідний метод, який потребує виконання 4-х кроків (рис. 4).

Крок 4.1. Для всіх прикладів i маркованої частини A_R^l навчальної вибірки формуємо двовимірну гістограму D (квадратну матрицю $\{D_{mk}\}$ $m, k = \overline{1,p}$) попарних збігів значень існуючих маркерів класів T і отриманих маркерів кластерів Z, $D_{mk} = D_{mk} + 1$, якщо $t_i^k = z_i^m$, де $m, k = \overline{1,p}$.

Крок 4.2. У циклах (за рядками та за стовпцями) обнуляємо елементи гістограми D_{mk}, залишаючи без зміни значення тих, які задовольняють правилу:

$$D_{mk}^* = \begin{cases} 0 \\ D_{mk}, \langle \max_m(D_{mk}) = \max_k(D_{mk}) | m = k | \max(D_{mk}) <> 0 \rangle \end{cases}. \quad (14)$$

Коригування гістограми D_{mk} виконується, поки в кожному рядку і стовпці залишиться тільки одне не нульове значення.

Крок 4.3. Перетворюємо скориговану гістограму D_{mk}^* в матрицю відповідності значень маркерів класів і кластерів $M_{(1,2)m}$ для всіх прикладів навчальної вибірки A_R^l.

Крок 4.4. Коригуємо вектор маркерів кластерів Z (3) за допомогою матриці відповідностей $M_{(1,2)m}$, формуємо класифікаційне рішення для маркованої частини навчальної вибірки у вигляді:

$$A_R^{l\,*} = \langle X, T, Z \rangle \left\langle \begin{bmatrix} x_{11} & \cdots & x_{1q} \\ \cdots & \cdots & \cdots \\ x_{l1} & \cdots & x_{lq} \end{bmatrix}, \begin{bmatrix} t_1^1 \\ \cdots \\ t_l^p \end{bmatrix}, \begin{bmatrix} z_1^1 \\ \cdots \\ z_l^p \end{bmatrix} \right\rangle, \quad (15)$$

і передаємо його експерту-соціологу для подальшої оцінки та інтерпретації.

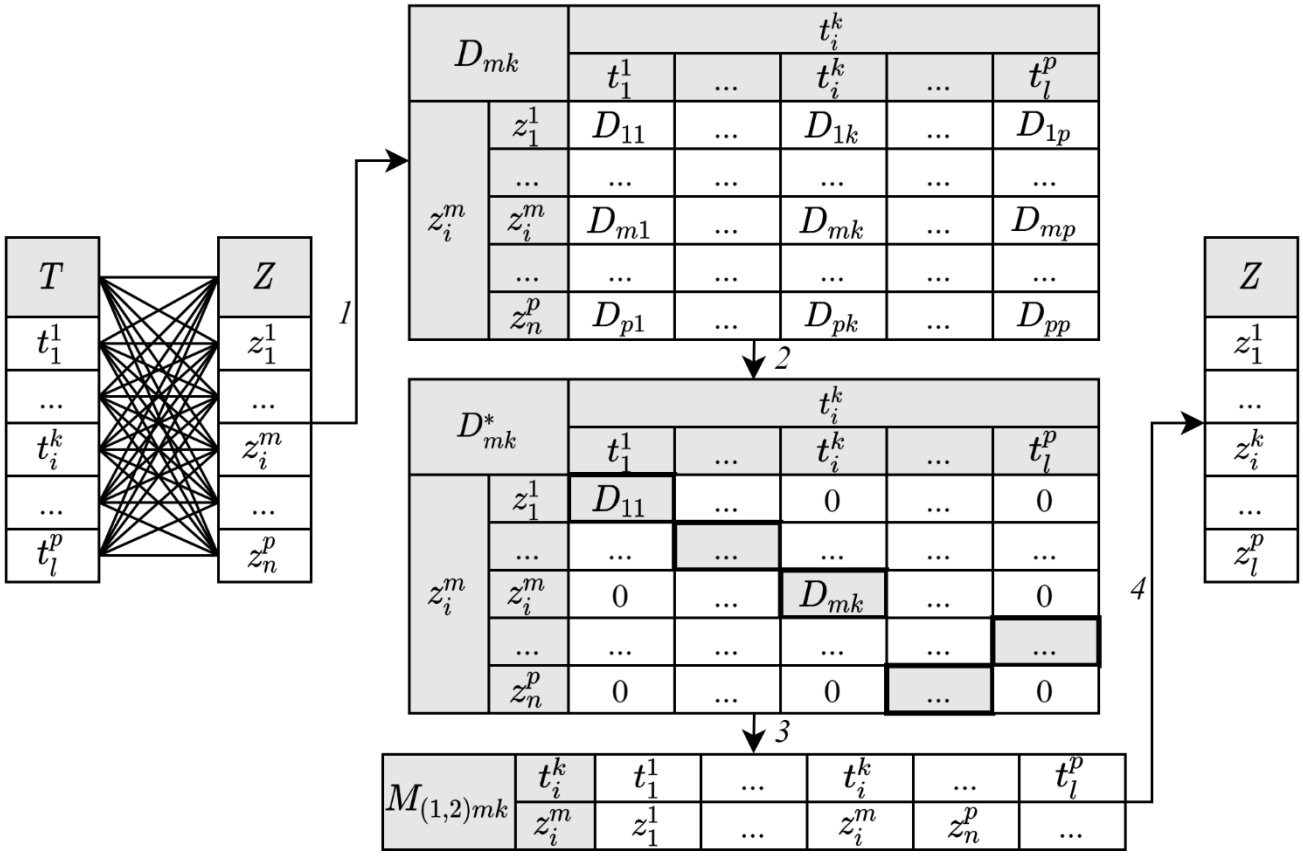


Рисунок 4 – Схема виконання кроків методу узгодження маркерів класів і кластерів

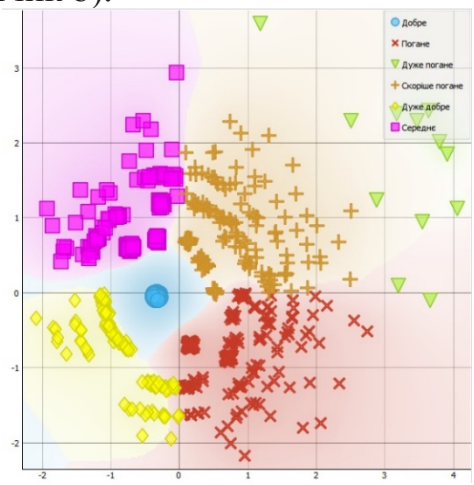
Етап 5. Остаточне *маркування агрегованих даних* A_R^{train} виконується експертом-соціологом для маркованої частини A_R^l та немаркованої частини A_R^u на основі рішення (8). Як правило навчальна вибірка A_R^l не є збалансованою, тому достовірність класифікації визначається за допомогою метрик *True Positive Rate* – TPR, яка показує відсоток прикладів позитивного класу, знайдених класифікатором, та *False Positive Rate* – *FPR,* яка показує відсоток прикладів негативного класу помилково віднесених класифікатором до позитивного класу:

$$TPR = \frac{TP}{TP + FN} \cdot 100\%, \qquad (16) \qquad FPR = \frac{FP}{FP + TN} \cdot 100\%, \qquad (17)$$

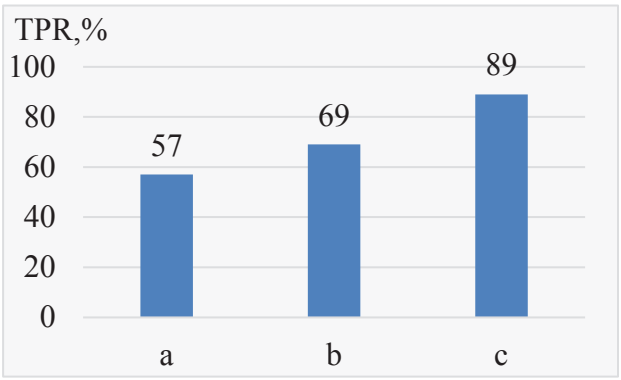
де TP – вірно класифіковані позитивні приклади; FN – позитивні приклади, класифіковані як негативні (помилка II роду), FP – негативні рішення, класифіковані як позитивні (помилка I роду); TN – істино негативні приклади, класифіковані як негативні

На рисунку 5,*а* показано карту подібності респондентів МСМ «Україна – стиль життя», яку побудовано за тим самим сценарієм, що і карти, показані на рисунках 2,*а* та 2,*б*, але для *шести* класів $\{t^k\}$, $k = \overline{1,6}$ в маркованій частині вибірки, а саме: ○ – «добре», × – «погане», ▽ – «дуже погане», + – «скоріше погане», ◇ – «дуже добре», □ – «середнє». При цьому маркери кластерів z^k, в немаркованій частині вибірки A_R^u отримано в автоматизованому режимі на основі самоорганізації мережі Кохонена (етап 2, див. рис. 3) та узгоджено з існуючими маркерами класів маркованої частини вибірки A_R^l. за розробленим методом (етап 4, див. рис. 3). На рисунку 5,*б* показані результати розрахунків достовірності класифікації за метрикою (16) (стовпчики a, b, c) для агрегованих даних A_R, MDS проекції, які відображено

відповідно на рисунках 2,*а*, 2,*б* та 5,*а*. При цьому необхідно зазначити, якщо час прийняття класифікаційних рішень за власним рішенням респондента прийняти за відносну «одиницю» (стовпчик a), тоді в автоматизованому режимі (стовпчик c) час знизився *вдвічі* в порівнянні з використанням ручної праці експерта-соціолога (стовпчик b).



а



б

Рисунок 5 – Результати реалізації методу навчання мережі Кохонена з частковим залученням вчителя респондентів МСМ за стилем життя: *а*) карта подібності; *б*) достовірність класифікації

Таким чином, використання методу узгодження маркерів класів та кластерів в процесі навчання мережі Кохонена за алгоритмом WTA з градієнтною оптимізацією дозволило автоматизувати прийняття класифікаційних рішень для *CC&CM* гетерогенних даних МСМ. Але достовірність класифікації в випадках близьких та незбалансованих класів складає менш ніж 90%. Тому для її підвищення запропоновано евристичне коригування вагових коефіцієнтів в процесі навчання мережі Кохонена.

В **третьому розділі удосконалено метод класифікації CC&CM гетерогенних даних МСМ** за рахунок узгодження існуючих маркерів класів та отриманих маркерів кластерів для подальшої оцінки пристосованості хромосом при евристичному коригуванні вагових коефіцієнтів в процесі навчання мережі Кохонена (рис. 6).

Етап 1. Ініціалізація мережі Кохонена та формування початкової популяції. Ініціалізація мережі Кохонена відбувається за описом Етапу 1 (див. рис. 3). Формування початкової популяції Pl полягає у створенні хромосом в кількості N_p. Кожна хромосома CH_h є вектором випадкових дійсних чисел, який співвідноситься із окремими ваговими коефіцієнтами та зміщеннями: $CH_h =< W, B >^h$, $h = \overline{1, N_p}$:

$$CH_h = \{w_{11}^h \dots, w_{1q}^h, \dots, w_{p1}^h, \dots, w_{pq}^h, b_1^h, \dots, b_p^h\}. \quad (18)$$

Якщо представити CH_h у вигляді множини генів $\{g_1, \dots, g_G\}$, де $G = p \times q + p$ – потужність CH_h, тоді популяція Pl виглядає наступним чином:

$$Pl = \begin{bmatrix} CH_1 \\ \cdots \\ CH_{N_p} \end{bmatrix} = \begin{bmatrix} g_{11} & \cdots & g_{1G} \\ \cdots & \cdots & \cdots \\ g_{N_p 1} & \cdots & g_{N_p G} \end{bmatrix}. \quad (19)$$

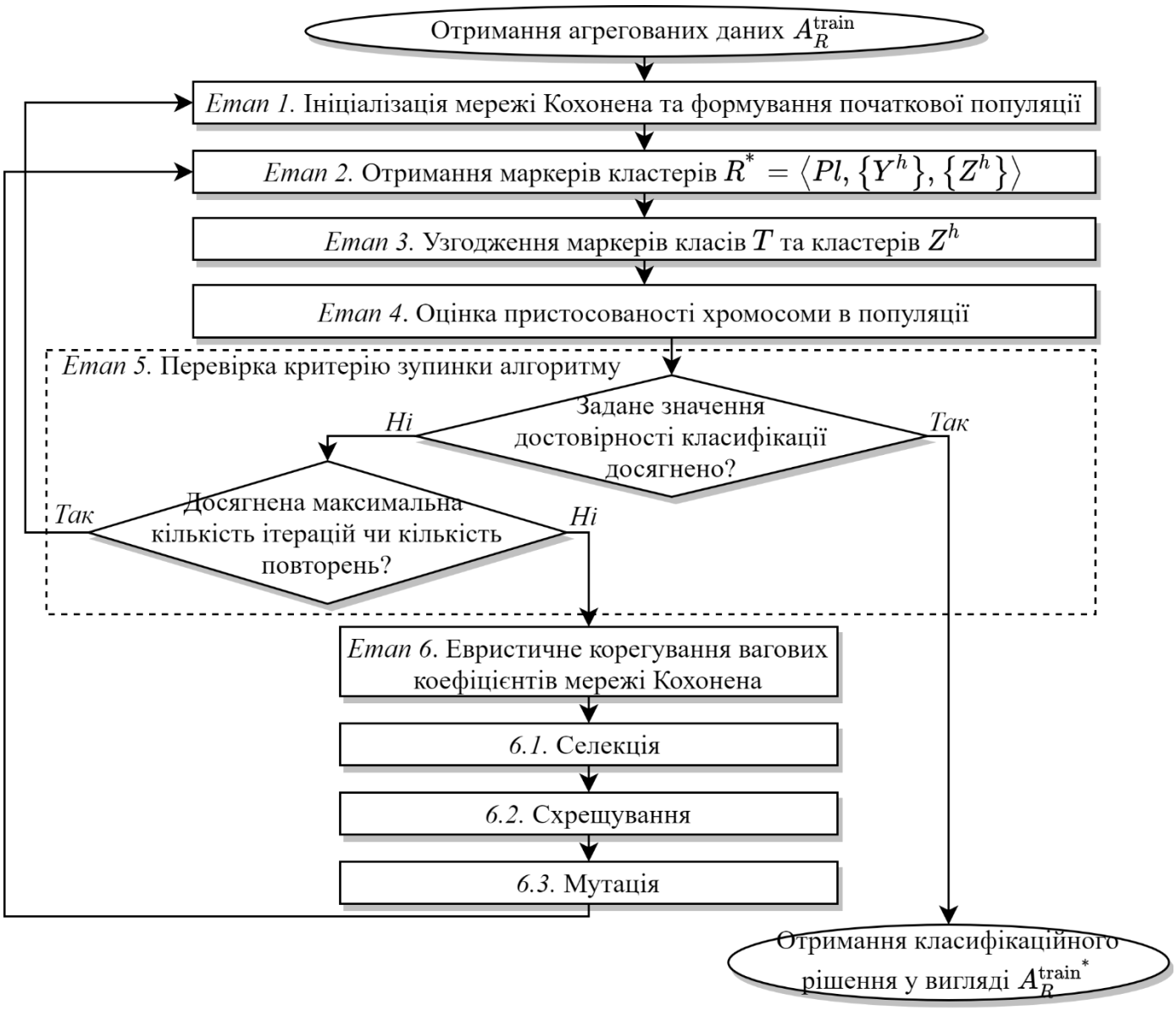


Рисунок 6 – Схема методу класифікації *CC&CM* гетерогенних даних МСМ

Етап 2. Отримання маркерів кластерів $R^* = \langle Pl, \{Y^h\}, \{Z^h\} \rangle$. Для отримання попередніх рішень у вигляді маркерів кластерів на вхід конкуруючого шару, який являє собою популяцію Pl (19), послідовно подаються значення векторів $\{x_{ki}\}$ з навчальної вибірки A_R^{train} та за евклідовою відстанню для кожної хромосоми CH_h розраховується відклик мережі у вигляді бінарного вектору $\{y_{kj}\}^h, k = \overline{1,n}, j = \overline{1,p}, h = \overline{1,N_p}$. Результатом етапу є вихідний вектор $\{z_{kj}\}^h$, значенню k-того елемента якого присвоюється порядковий номер нейрона-переможця j бінарного вектору $\{y_{kj}\}^h$. Таким чином, для популяції Pl з врахуванням кожної хромосоми $CH_h, h = \overline{1,N_p}$ отримаємо множину попередніх рішень в наступному вигляді:

$$R^* = \langle Pl, \{Y^h\}, \{Z^h\} \rangle = \langle \begin{bmatrix} x_{11} & \cdots & x_{1q} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nq} \end{bmatrix}, \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \cdots & \cdots & \cdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix}^h, \begin{bmatrix} z_{11} \\ \cdots \\ z_{np} \end{bmatrix}^h \rangle, \quad (20)$$

де X – множина ознак респондентів R_i, Y^h – бінарна матриця відкликів конкуруючого шару, Z^h– множина маркерів кластерів.

Етап 3. Узгодження маркерів класів T та кластерів Z^h виконується для кожної хромосоми CH_h $h = \overline{1, N_p}$ згідно розробленого методу (рис. 4), отримуємо:

$$R^{**} = \langle Pl, X, T, \{Z^h\} \rangle = \langle \begin{bmatrix} x_{11} & \dots & x_{1q} \\ \dots & \dots & \dots \\ x_{l1} & \dots & x_{lq} \end{bmatrix}, \begin{bmatrix} t_1^1 \\ \dots \\ t_l^p \end{bmatrix}, \begin{bmatrix} z_1^1 \\ \dots \\ z_l^p \end{bmatrix}^h \rangle. \quad (21)$$

Етап 4. Оцінка пристосованості хромосоми в популяції. У циклі на основі існуючих маркерів класів T і узгоджених маркерів кластерів Z^h, розраховуємо достовірність класифікації TPR^h за метрикою (16) для кожної хромосоми CH_h.

Етап 5. Перевірка критерію зупинки алгоритму. Критеріями зупинки роботи алгоритму є: *вдале завершення* – досягнення необхідного рівня достовірності класифікації; *невдале завершення* – перевищення максимальної кількості ітерацій або повторень значення достовірності класифікації, яке означає виродження популяції та вимагає переходу до повторної ініціалізації мережі Кохонена. Якщо умови зупинки не виконуються переходимо до шостого етапу.

Етап 6. Евристичне коригування вагових коефіцієнтів мережі Кохонена відбувається за виконанням класичних процедур *селекції* (*selection*), *схрещування* (*crossover*) та *мутації (mutation)* хромосом CH_h за отриманою на Етапі 5 оцінкою їх пристосованості. Для виконання селекції сортуємо популяції в порядку зменшення оцінки пристосованості TPR^h. Із відсортованої послідовності обираємо дві перші (CH_1^w (*winner*) і CH_2^{vw} (*vice winner*)) та дві останні ($CH_{N_p-1}^{vl}$ (*vice loser*) і $CH_{N_p}^{l}$ (*loser*)) хромосоми. Переможці схрещуються для отримання двох «пристосованих» хромосом з метою подальшої заміни «лузерних» хромосом. Для схрещування, визначають випадкову точку кросинговеру, яка розділяє хромосому на дві частини та проводять обмін генами хромосом (рис. 7).

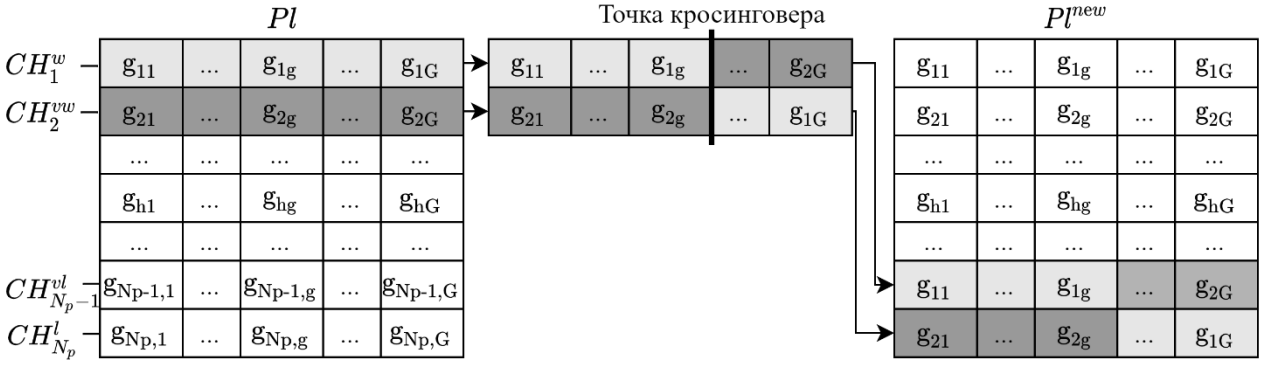


Рисунок 7 – Приклад оператора схрещування

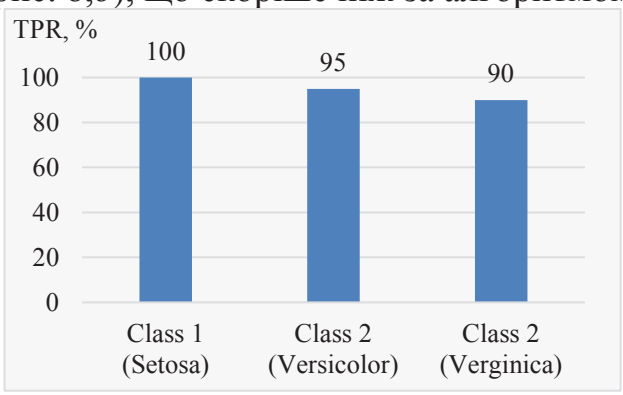
Операція мутації є допоміжною при створенні нової хромосоми та виконується додаванням до кожного гену хромосоми CH_h випадкового числа, розподіленого за рівномірним законом:

$$CH_h^{new} = CH_h + rand(-0{,}05\max(CH_h), +0{,}05\max(CH_h)). \quad (22)$$

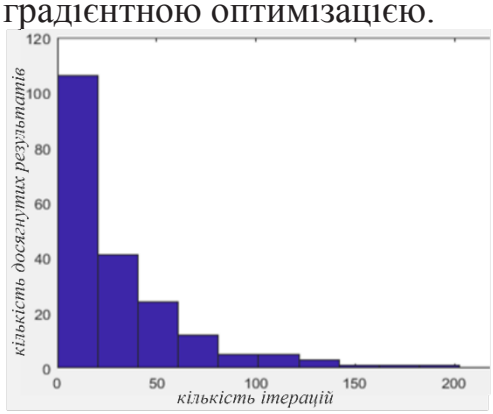
Із нових хромосом формується нова популяція Pl^{new}, яку в якості вхідних даних передаємо для виконання дій етапу 2.

Результати апробації методу евристичного коригування вагових коефіцієнтів при навчанні мережі Кохонена для прикладу класифікації даних про «іриси Фішера» з класами 2 та 3, які перетинаються, показано на рис. 8.

Аналіз наведеної діаграми навчання засвідчує, що всі 200 випробувань навчання мережі Кохонена набули вдалого завершення з достовірністю класифікації більш ніж 95% та 90% для класів, що перетинаються. При цьому у більшості випробуваній (176) для навчання мережі Кохонена було потрібно менше ніж 50 епох (рис. 8,*б*), що скоріше ніж за алгоритмом WTA з градієнтною оптимізацією.



а



б

Рисунок 8 – Результат апробації методу евристичного коригування вагових коефіцієнтів мережі Кохонена: *а*) TPR, *б*) діаграма навчання

Таким чином, **отримав подальший розвиток метод евристичного коригування вагових коефіцієнтів в процесі навчання нейронної мережі Кохонена**, що дозволило удосконалити метод класифікації *CC&CM* гетерогенних даних МСМ (див. рис. 6, етапи 4-6).

Данні про апробацію удосконаленого методу класифікації *CC&CM* гетерогенних даних респондентів МСМ «Україна - стиль життя» (див. рис. 2 та 5) наведено в таблиці 1. Порівняльні результати, щодо часу та достовірності прийняття класифікаційних рішень на основі мережі Кохонена з коригуванням вагових коефіцієнтів за алгоритмом WTA та з евристичним коригуванням (*Heuristic adjustment* – HA) засвічують про перевагу удосконаленого методу класифікації на навчальній (*training)* та тестовій (*testing)* вибірках гетерогенних даних респондентів МСМ.

Таблиця 1. Порівняння результатів класифікації *CC&CM* гетерогенних даних МСМ

Показники	WTA		HA	
	Training	Testing	Training	Testing
Достовірність класифікації (TPR), %	88,96 %	81,08 %	94,99 %	86,06 %
Час роботи, с	1541		1217	

В **четвертому розділі** розроблено ІС МСМ (рис. 9). Підсистема підготовки вихідних даних ІС МСМ побудована за принципом ETL-систем, що містить засоби:

– вилучення (*Extraction*) деталізованих даних різних форматів з різнотипних інформаційних джерел про об'єкт управління та їх зберігання з врахуванням розробленої моделі деталізованих даних;

– агрегування і перетворення (*Transformation*) даних МСМ в єдиний формат на основі розробленої моделі агрегованих даних;

– завантаження (*Loading*) агрегованих даних (навчальних вибірок) в сховище для подальшого використання в якості навчальної вибірки.

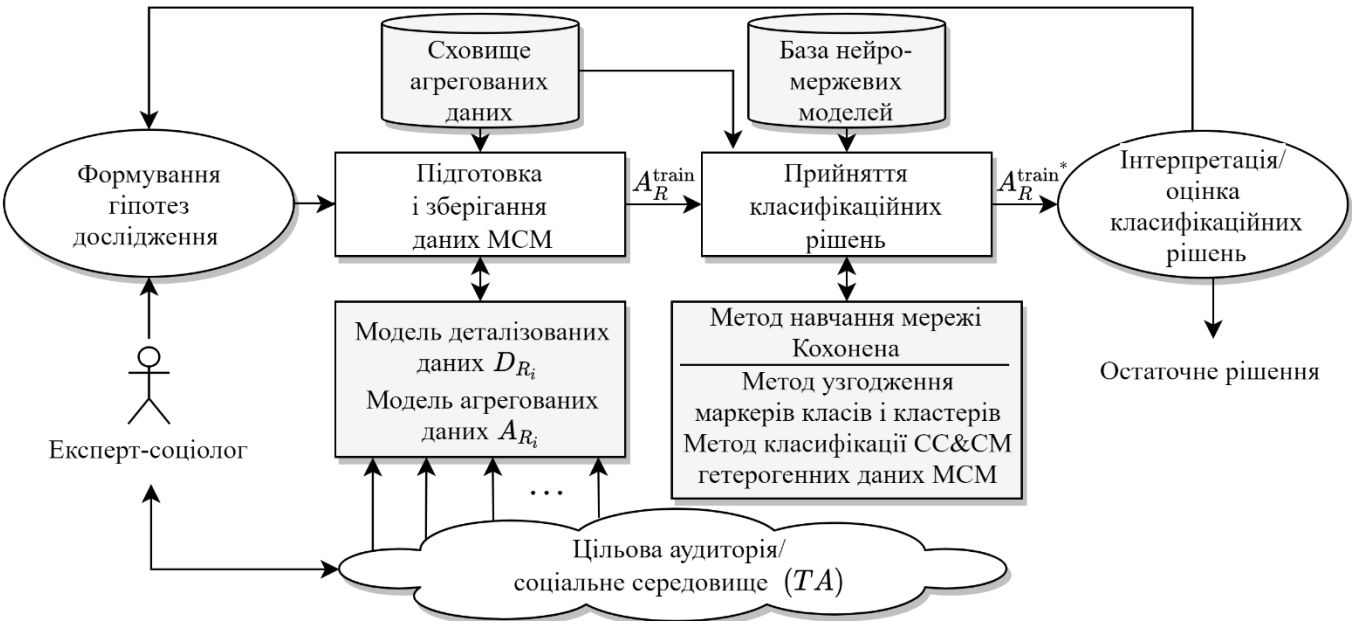
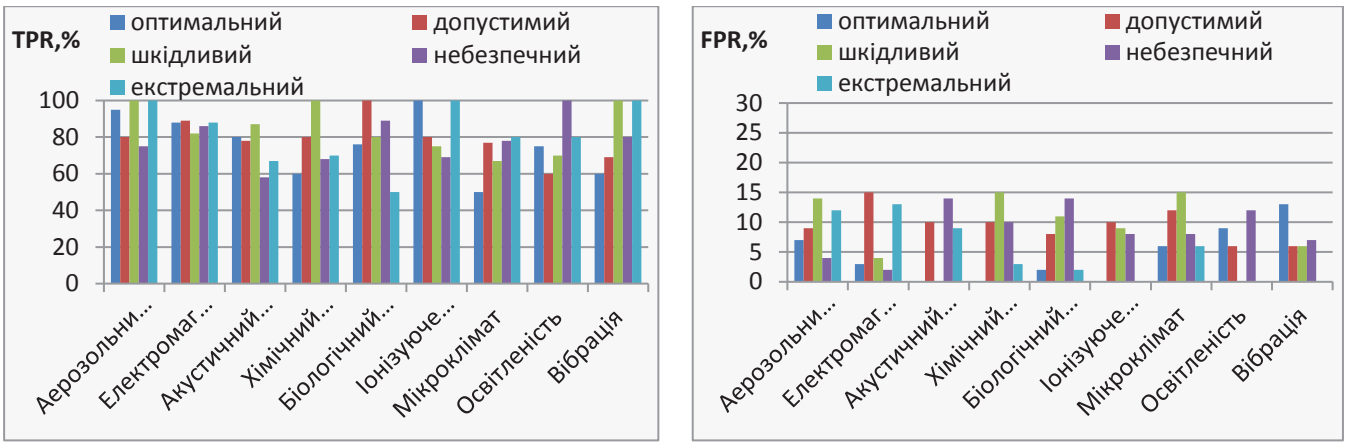


Рисунок 9 – Структура ІС МСМ, з врахуванням запропонованих моделей та методів

При реалізації блока навчання мережі Кохонена враховано розроблений метод класифікації *CC&CM* гетерогенних даних МСМ за рахунок узгодження існуючих маркерів класів та отриманих маркерів кластерів для подальшої оцінки пристосованості хромосом при евристичному коригуванні вагових коефіцієнтів в процесі навчання мережі Кохонена. Отримані матриці вагових коефіцієнтів завантажуються в базу нейромережевих моделей та далі спільно з агрегованими даними використовуються для уточнення гіпотези дослідження про склад цільової аудиторії, підтримки прийняття класифікаційних рішень для підтвердження або спростування уточненої гіпотези дослідження.

Розроблену ІС МСМ використано для класифікації даних «умови праці» працівників підприємства ТОВ «Сервіс-Грейн» при проведенні комплексної експертизи факторів які впливають на стан здоров'я цільової аудиторії. Умови праці залежать від кількісно-якісних ознак, які задають рівні аерозольного, електромагнітного, акустичного, хімічного і біологічного впливу, іонізуючого випромінювання, мікроклімату, освітленості і вібрації. В результаті аналізу даних МСМ відповідно до запропонованої шкали, яка залежить від рівня гранично допустимої концентрації досліджуваного фактору, експерт-соціолог оцінює стан умов праці як: «*оптимальний*», «*допустимий*», «*шкідливий*», «*небезпечний*», «*екстремальний*». Навчальна вибірка складалася із 1200 прикладів. В ході дослідження експерт-соціолог має можливість порівняти та відкоригувати своє рішення у відповідності до автоматизованого рішення ІС МСМ. Достовірність класифікації визначалась за метриками (16) – *TPR* та (17) – *FPR*.

Аналіз розрахунків значень помилок до і після проведення коригування показує підвищення відносної частки *TPR* в середньому на 20% і зниження на 50% *FPR* для всіх груп умов праці. При цьому після проведення коригування рішень експерта-соціолога, оцінки *TPR* розташовуються більш рівномірно для всіх досліджуваних рівнів впливу з групи умов праці (рис. 10,*а*). А зниження відносної частки істинно негативних випадків (*FPR*) особливо помітно для рівнів хімічного і біологічного впливу (рис. 10, *б*).



а *б*

Рисунок 10 – Результати класифікації умов праці: *а*) TPR; *б*) FPR

ВИСНОВКИ

Дисертаційна робота містить раніше не захищені наукові положення та одержані автором нові науково-обґрунтовані результати, які полягають в розробці та дослідженні моделей представлення *CC&CM* гетерогенних даних МСМ та методів їх класифікації, з метою зниження часу на прийняття класифікаційних рішень без втрати рівня їх достовірності в ІС МСМ.

1. Проведений аналіз структури сучасних ІС МСМ та процесу прийняття класифікаційних рішень на їх основі дозволив встановити, що для зниження часу класифікації без втрати рівня її достовірності потребують удосконалення підсистеми попередньої обробки та прийняття класифікаційних рішень. Встановлено, що гетерогенні дані МСМ є слабко структурованими, тому використання глибинних нейронних мереж зі вчителем для класифікації таких даних потребує їх тривалої попередньої обробки, що значно знижує необхідну в умовах моніторингу швидкість або достовірність прийняття класифікаційних рішень.

2. За допомогою побудованих карт подібності респондентів доведено, що слабко структуровані гетерогенні дані МСМ є слабко маркованими тому, що в умовах їхнього періодичного оновлення та тривалості класифікації в ручному та автоматизованому режимах в типовій ІС МСМ складно отримати достатню кількість класифікаційних рішень у вигляді попередньо розмічених даних.

3. Для удосконалення підсистеми попередньої обробки в ІС МСМ розроблено моделі представлення *CC&CM* гетерогенних даних МСМ в просторах властивостей та ознак з урахуванням значень, типів, форматів, джерел, оцінок якості та процедури агрегування/трансформації властивостей деталізованих даних. Це дозволило формалізувати створення навчальної та тестової вибірок для подальшої класифікації з частковим залученням вчителя.

4. Для отримання додаткових маркерів класів в немаркованій частині навчальної вибірки розроблено метод узгодження маркерів класів та кластерів при навчанні мережі Кохонена, який заснований на побудові двовимірної гістограми попарних збігів значень маркерів класів і кластерів з її подальшим коригуванням за рядками та за стовпцями у відповідності до розробленого правила. Апробація

методу показала, що час прийняття класифікаційних рішень в автоматизованому режимі знизився *вдвічі* в порівнянні з ручним режимом.

5. Розроблено метод евристичного коригування вагових коефіцієнтів в процесі навчання нейронної мережі Кохонена з узгодженням маркерів класів та кластерів для розрахунку оцінки пристосованості хромосоми в популяції генетичного алгоритму для кожного прикладу із навчальної вибірки. Аналіз наведених діаграм навчання мережі Кохонена для низки випробувань на вибірках з класами, які перетинаються, засвідчує про отримання достовірності класифікації більш ніж 90-95% за менше ніж 50 епох навчання, що скоріше на 50% ніж за алгоритмом WTA з градієнтною оптимізацією.

6. Розроблено метод класифікації на базі узгодження існуючих маркерів класів та отриманих маркерів кластерів для подальшої оцінки пристосованості хромосом при евристичному коригуванні вагових коефіцієнтів в процесі навчання мережі Кохонена. Порівняльні результати щодо часу та достовірності прийняття класифікаційних рішень на основі мережі Кохонена з коригуванням вагових коефіцієнтів за алгоритмом WTA та з евристичним коригуванням засвічують про переваги удосконаленого методу класифікації на навчальній та тестовій вибірках гетерогенних даних респондентів МСМ «Україна - стиль життя», а саме збільшення достовірності на 5% та зменшення часу на 21%.

7. Практичне використання моделей представлення *CC&CM* гетерогенних даних МСМ та методів їх класифікації дозволило розробити ІС МСМ, до складу якої входять оновлені підсистеми попередньої обробки та прийняття класифікаційних рішень. Апробація ІС МСМ на даних про стиль життя та умови праці респондентів показала збільшення достовірності та зменшення часу позитивні, що підтверджено відповідними актами впровадження.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації:

1. Арсирий, Е.А. Автоматизация представления и извлечения декларативных знаний в цифровом макете предприятия при инициализации проектов охраны труда / Е. А. Арсирий, С. Г. Антощук, О. С. Маникаева // Электротехнические и компьютерные системы – 19(95), 2015. С. 266 – 271.

Стаття опублікована у виданні України, яке включене до міжнародних наукометричних баз eLIBRARY, ULRICHSWEB, FreeFullPDF.

2. Маникаева, О. С. Машинное обучение самоорганизующегося слоя Кохонена в системах нейросетевого распознавания образов по статистической информации / О. С. Маникаева, Е. А. Арсирий, А. П. Василевская // Системні технології. Регіональний міжвузівський збірник наукових праць. – Вип. 5 (100). – Дніпропетровськ, 2015. С. 93 – 105.

3. Арсирий, Е. А. Разработка подсистемы поддержки принятия решений в системах нейросетевого распознавания образов по статистической информации / Е. А. Арсирий, О. С. Маникаева, А. П. Василевская // Восточно-Европейский журнал передовых технологий. –2015 – №6/4(78). – С. 4 –12.

Стаття опублікована у виданні України, яке включене до міжнародних наукометричних баз Scopus eLIBRARY, ULRICHSWEB, FreeFullPDF.

4. Arsirii O. O. Models and methods of intellectual analysis for medical-sociological monitoring's data based on the neural network with a competitive layer / O. O. Arsirii, O. S. Manikaeva // Applied Aspects of Information Technology. – Odessa, 2019. – Vol. 2 № 3. – P. 173-185

Стаття опублікована у виданні України, яке включене до міжнародних наукометричних баз eLIBRARY, ULRICHSWEB, FreeFullPDF

5. Arsirii O. Heuristic models and methods for application of the Kohonen neural network in the intellectual system of Medical-Sociological Monitoring / O. Arsirii, O. Manikaeva, A. Nikolenko, O. Babilunha // Herald of Advanced Information Technology. – 2020, Vol. 3, № 1. – P. 395-405

Стаття опублікована у виданні України, яке включене до міжнародних наукометричних баз eLIBRARY, ULRICHSWEB, FreeFullPDF

Наукові праці апробаційного характеру:

6. Arsirii O. O. Automation of the preparation process weakly-structured multi-dimensional data of sociological surveys in the data mining system / O. O. Arsirii, O. Yu Babilunha, O. S. Manikaeva, O. I. Rudenko // Herald of Advanced Information Technology. – 2018. – Vol. 1, №1. P. 11–20.

Стаття опублікована у виданні України, яке включене до міжнародних наукометричних баз eLIBRARY, ULRICHSWEB, FreeFullPDF

7. Arsirii O. Intellectual information technology of Analysis of weakly-structured multi-dimensional data of sociological research / Arsirii O., Antoshchuk S., Babilunha O., Manikaeva O., Nikolenko // Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2019. Advances in Intelligent Systems and Computing. – 2019. – Vol. 1020. Springer, Cham. – P. 242-258.

Стаття опублікована у виданні України, яке включене до міжнародних наукометричних баз Scopus eLIBRARY, ULRICHSWEB, FreeFullPDF

8. Манікаєва О. С. Використання нейронних мереж для оцінки стану умов праці / О. С. Манікаєва, А. Ю. Москалюк // V Міжнародна науково-практична конференція магістрантів, аспірантів та науковців «Управління проектами: інновації, нелінійність, синергетика». – Одеса: ОДБА, 2014. – С. 147 – 149.

9. Арсирий Е. А. Автоматизация представления и извлечения декларативных знаний в цифровом макете предприятия при инициализации проектов охраны труда / Е. А. Арсирий, О. С. Маникаева, А. Ю. Москалюк // ІІІ українсько-німецька конференція «Інформатика. Культура. Техніка» – Одеса: ОНПУ, 2015. – С. 13 – 14.

10. Арсирий Е. А. Сравнение методов извлечения декларативных знаний в цифровом макете предприятия при инициализации проектов охраны труда / Е. А. Арсирий, О. С. Маникаева // IV міжнародна науково-практична конференція «Інформаційні системи та технології». – Харків: ХНУРЕ , 2015. – С. 26 – 27.

11. Маникаева О. С. Подсистема поддержки принятия решений в системе мониторинга функционального состояния сложного объекта труда / Е. А. Арсирий, О. С. Маникаева // XVII міжнародна науково-практична конференція «Сучасні інформаційні та електронні технології». – Одеса: ОНПУ, 2016. – С. 50 – 51.

12. Арсирий Е. А. Проектирование нейронной сети ART-1 для принятия классификационных решений при инициализации проектов охраны труда // Е. А. Арсирий, О. С. Маникаева // VI міжнародна науково-практична конференція

«Інформаційні управляючі системи та технології» – Одеса: ОНМУ, 2017. – С. 39 – 42.

13. Manikaeva O. Development of optimization method to train Kohonen neural network based on genetic algorithm // Manikaeva O., Federico Barber Sanchís // V українсько-німецька конференція «Інформатика. Культура. Техніка». – Одеса: ОНПУ, 2017. – С. 13 – 14.

АНОТАЦІЯ

Манікаєва О. С. Методи класифікації гетерогенних даних в інтелектуальних системах медико-соціологічного моніторингу. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 – «Системи та засоби штучного інтелекту». – Одеський національний політехнічний університет МОН України, Одеса, 2020.

Дисертація присвячена вирішенню науково-практичної задачі створення інтелектуальних систем медико-соціологічного моніторингу (МСМ) на основі удосконалених моделей представлення гетерогенних даних МСМ та розроблених методів їх класифікації за допомогою навчання мережі Кохонена. В роботі удосконалено моделі представлення слабко структурованих та слабко маркованих гетерогенних даних МСМ в просторах властивостей та ознак з урахуванням значень, типів, форматів, джерел, оцінок якості та процедури агрегування/трансформації властивостей деталізованих даних. На основі цих моделей розроблено метод узгодження маркерів класів та кластерів при навчанні мережі Кохонена з частковим залученням вчителя, який заснований на побудові двовимірної гістограми попарних збігів значень маркерів класів і кластерів з її подальшим коригуванням за рядками та за стовпцями у відповідності до розробленого правила. Метод дозволяє отримати додаткові маркери класів в немаркованій частині навчальної вибірки. На базі методу узгодження маркерів класів та кластерів для оцінки пристосованості хромосоми в популяції генетичного алгоритму для кожного прикладу із навчальної вибірки запропоновано метод евристичного коригування вагових коефіцієнтів в процесі навчання нейронної мережі Кохонена. Використання такого коригування вагових коефіцієнтів дозволило зменшити час навчання мережі Кохонена без втрати рівня достовірності класифікації. Удосконалено метод класифікації слабко структурованих та слабко маркованих гетерогенних даних МСМ за рахунок використання розроблених методів узгодження маркерів класів та кластерів та евристичного коригування вагових коефіцієнтів в процесі навчання мережі Кохонена.

Ключові слова: інтелектуальні системи, гетерогенні дані, слабко структуровані дані, слабко марковані дані, нейронна мережа Кохонена, генетичний алгоритм.

ABSTRACT

Manikaeva O. Classification methods of heterogeneous data in intellectual systems of medical and social monitoring. – Manuscript

The thesis for a candidate's degree of technical sciences by specialty 05.13.23 – "Systems and means of artificial intelligence". – Odessa National Polytechnic University of the Ministry of Education and Science of Ukraine, Odessa, 2020.

The thesis is devoted to the solution of the scientific and practical problem of creation of intelligent systems (IS) of medical and sociological monitoring (MSM) on the basis of the improved models of representation of heterogeneous MSM's data and the developed methods of their classification by means of Kohonen network training. In the research improve the models of representation of weakly structured and weakly labeled heterogeneous MSM's data in the spaces of properties and features, taking into account values, types, formats, sources, quality assessments and procedures of aggregation / transformation of properties of detailed data. Based on these models, a method of matching classes and clusters markers in learning of Kohonen network with partial teacher involvement was developed, which is based on constructing a two-dimensional histogram of pairwise matches of classes and clusters markers values with its subsequent intersecting by rows and by columns to the developed rule. The method allows to obtain additional class markers in the unlabeled part of the training sample. Based on the method of matching classes and clusters markers to assess the suitability of chromosomes in the population of the genetic algorithm for each example from the training sample, a method of heuristic weight adjustment in the learning process of the Kohonen neural network is proposed. The use of such an adjustment of the weights allowed to reduce the training time of the Kohonen network without losing the level of reliability of the classification. The method of classification of weakly structured and weakly labeled heterogeneous MSM's data has been improved due to the use of developed methods of matching class and cluster markers and heuristic adjustment in the process of learning the Kohonen network.

Keywords: intelligent systems, heterogeneous data, weakly structured data, weakly labeled data, Kohonen neural network, genetic algorithm.

АННОТАЦИЯ

Маникаева О. С. Методы классификации гетерогенных данных в интеллектуальных системах медико-социологического мониторинга. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 - «Системы и средства искусственного интеллекта». – Одесский национальный политехнический университет МОН Украины, Одесса, 2020.

Диссертация посвящена решению научно-практической задачи разработки моделей представления слабо структурированных и слабо маркированных гетерогенных данных медико-социологического мониторинга и методов их классификации, с целью снижения времени на принятие классификационных решений без потери уровня их достоверности в интеллектуальных системах (ИС) медико-социологического мониторинга (МСМ). Проведенный анализ структуры современных ИС МСМ и процесса принятия классификационных решений на их основе позволил установить, что для снижения времени классификации без потери уровня их достоверности усовершенствования требуют две подсистемы ИС МСМ: предварительной обработки и принятия классификационных решений. Установлено, что гетерогенные данные МСМ являются слабо структурированными (*СС*), поэтому использование глубинных нейронных сетей с учителем для классификации таких данных требует их длительной предварительной обработки, что значительно снижает необходимую в условиях мониторинга скорость или достоверность

принятия классификационных решений. С помощью построенных карт сходства респондентов доказано, что *CC* гетерогенные данные МСМ являются еще и слабо маркированными (*CM*) потому, что в условиях их периодического обновления, продолжительность классификации в ручном и автоматизированном режимах не дает возможности получить достаточное количество классификационных решений в виде предварительно размеченных данных.

Для усовершенствования подсистемы предварительной обработки в ИС МСМ разработаны модели представления *CC&CM* гетерогенных данных МСМ в пространствах свойств и признаков с учетом значений, типов, форматов, источников, оценок качества и процедуры агрегирования/трансформации свойств детализированных данных.

Для получения дополнительных маркеров классов в немаркированной части обучающей выборки разработан метод согласования маркеров классов и кластеров для обучения сети Кохонена с частичным привлечением учителя, который основан на построении двумерной гистограммы попарных совпадений значений маркеров классов и кластеров с ее последующей построчно-постолбцовой корректировкой в соответствии с разработанным правилом. Апробация метода показала, что время принятия классификационных решений в автоматизированном режиме снизилось в два раза по сравнению с использованием ручного труда эксперта-социолога.

Для уменьшения времени классификации *CC&CM* гетерогенных данных МСМ без потери уровня ее достоверности разработан метод эвристической корректировки весовых коэффициентов в процессе обучения нейронной сети Кохонена на основе использования метода согласования маркеров классов и кластеров для расчета оценки приспособленности хромосомы в популяции на основе генетического алгоритма для каждого примера из обучающей выборки. Анализ приведенных диаграмм обучения для ряда испытаний обучения сети Кохонена на выборках с пересекающимися классами свидетельствует о получении достоверности классификации более чем 90-95% за менее чем 50 эпох обучения, что значительно быстрее, чем с использованием алгоритма WTA.

Для усовершенствования подсистемы принятия классификационных решений в ИС МСМ разработан метод классификации на базе согласования существующих маркеров классов и полученных маркеров кластеров для дальнейшей оценки приспособленности хромосом на основе эвристической корректировки весовых коэффициентов в процессе обучения сети Кохонена. Сравнительные результаты о времени и достоверности принятия классификационных решений на основе сети Кохонена с корректировкой весовых коэффициентов по алгоритму WTA и с эвристической корректировкой свидетельствуют о преимуществах усовершенствованного метода классификации на обучающей и тестовой выборках гетерогенных данных респондентов МСМ «Украина - стиль жизни».

Апробация ИС МСМ на данных о стиле жизни и условиях труда респондентов показала положительные результаты, которые подтверждены соответствующими актами внедрения.

Ключевые слова: интеллектуальные системы, гетерогенные данные, слабо структурированные данные, слабо маркированные данные, генетический алгоритм.