

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”

Годич Олесь Васильович

УДК 004.853+004.855.5

**ІНДУКТИВНІ МЕТОДИ ТА АЛГОРИТМИ САМООРГАНІЗАЦІЇ
МОДЕЛЕЙ ДАНИХ НА ОСНОВІ КАРТ КОХОНЕНА**

01.05.03 — математичне та програмне забезпечення обчислювальних
машин і систем

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Львів — 2010

Дисертацією є рукопис.

Робота виконана в Львівському національному університеті імені Івана Франка Міністерства освіти і науки України.

Науковий керівник: кандидат фізико-математичних наук, доцент
Щербина Юрій Миколайович,
Львівський національний університет
імені Івана Франка, доцент кафедри дискретного
аналізу та інтелектуальних систем.

Офіційні опоненти: доктор технічних наук, доцент
Романишин Юрій Михайлович,
Національний університет “Львівська політехніка”,
професор кафедри електронних засобів
інформаційно-комп’ютерних технологій;
кандидат технічних наук, доцент
Ковалюк Тетяна Володимирівна,
Національний технічний університет України
“Київський політехнічний інститут”,
доцент кафедри автоматизованих систем обробки
інформації та управління.

Захист відбудеться «___» _____ 2010 р. о ___ годині на засіданні спеціалізованої вченої ради Д 35.052.05 Національного університету “Львівська політехніка” за адресою: 79013, м. Львів, вул. С. Бандери, 12.

З дисертацією можна ознайомитись у бібліотеці Національного університету “Львівська політехніка” за адресою: 79013, м. Львів, вул. Професорська, 1.

Автореферат розісланий «___» _____ 2010 р.

Учений секретар
спеціалізованої вченої ради
доктор технічних наук, професор



Р. А. Бунь

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Із розвитком і поширенням обчислювальної техніки в різних галузях діяльності людини об'єми даних, які зберігаються у файлах та базах даних, збільшуються високими темпами. Водночас користувачі, які працюють із цими даними, потребують кращих засобів отримання з них знань. Наприклад, за базою даних пацієнтів, яка містить різноманітні результати медичних аналізів, недостатньо отримати перелік тих пацієнтів, у яких невідповідний рівень гемоглобіну, що легко досягається структурованою мовою запитів. Сьогодення вимагає більш складного аналізу даних. Вирішенням таких проблем займається напрямок аналізу даних із назвою *виявлення знань*¹ (knowledge discovery). Одним із етапів виявлення знань є *видобування даних* (U. Fayyad). У працях (D. Pyle, E. Triantaphyllou, G. Felici, M. Dunham) поняття *видобування даних* пов'язують із набором підходів та інструментів для пошуку прихованої інформації у даних. Причому, дані можуть бути структурованими (бази даних) і неструктурованими (графічні зображення).

Традиційні засоби запитів баз даних передбачають добре означену мову запитів. Формуючи запит, дослідник повинен чітко знати, що саме його цікавить, і результатом запиту є завжди підмножина даних (можливо агрегованих). Натомість, у *видобуванні даних* переважно не передбачено існування структурованої мови запитів, дослідник може не знати, чого очікувати від аналізу даних, результат пошуку переважно не є підмножиною досліджуваних даних. Вважається, що на теперішній час рівень розвитку теорії *видобування даних* є на рівні розвитку засобів побудови запитів до баз даних початку 1970-х років (M. Dunham). Усі існуючі методи і алгоритми *видобування даних* спрямовані на знаходження адекватних моделей даних. Із підвищенням швидкодії обчислювальної техніки зростає актуальність використання саме індуктивних підходів аналізу і побудови моделей даних.

Ця дисертація присвячена методам дослідницького аналізу даних і побудові описових моделей даних із використанням процесів самоорганізації, які базуються на теорії карт Кохонена. Розроблені методи застосовані до аналізу структурованих і неструктурованих даних. Важливе місце у дисертації займає дослідження та розроблення методів візуалізації даних відповідно до концепції візуального пошуку інформації (B. Shneiderman). Вирішено завдання організації ефективних обчислень при програмній реалізації методів аналізу даних, які враховують архітектурні особливості сучасної обчислювальної техніки і тенденції її розвитку.

Зазначимо, що теорія самоорганізованих карт Кохонена мало використовується у працях вітчизняних науковців. Водночас, у світовій науці та

¹В україномовних джерелах все частіше зустрічається термін *інтелектуальний аналіз даних*.

практиці вона належить до провідних напрямків досліджень у контексті видобування даних (Т. Kohonen, S. Haykin, R. Begg, J. Kamruzzaman, S. K. Halgamuge, L. Wang).

Зв'язок роботи з науковими програмами, планами, темами.

Тема дисертації узгоджена із загальним планом науково-дослідних робіт кафедри дискретного аналізу та інтелектуальних систем Львівського національного університету імені Івана Франка: “Нелінійний аналіз і методи в задачах оптимізації” (I кв.2000р.– IV кв.2002р., № держ. реєстрації 0103U001084), “Ітераційні методи з надлінійною швидкістю збіжності розв’язування нелінійних операторних рівнянь і задач на екстремум” (I кв.2003р.– IV кв.2005р., № держ. реєстрації 0104U002135), “Чисельний аналіз лінійних динамічних систем. Алгоритмізація процесів прийняття рішень” (I кв.2004р.– IV кв.2006р., № держ. реєстрації 0104U004661), “Розробка Лі-алгебраїчних, ітеративних методів для динамічних систем, задач оптимізації та навчання нейронних мереж” (I кв.2005р.– IV кв.2007р., № держ. реєстрації 0105U002232), “Математичні методи дослідження динамічних та адаптивних систем і задач оптимізації” (I кв.2008р.– IV кв.2009р., № держ. реєстрації 0108U000749), “Математичні методи дослідження нелінійних динамічних систем, індуктивних методів моделювання даних і задач оптимізації” (I кв.2010р.– IV кв.2011р., № держ. реєстрації 0110U001373). Автором розроблено індуктивні методи та алгоритми самоорганізації моделей даних на основі карт Кохонена.

Мета і завдання дослідження. Метою дисертаційного дослідження є розвиток індуктивних методів аналізу даних на основі самоорганізованих карт Кохонена і розроблення ефективного математичного та програмного забезпечення комп’ютерної системи аналізу даних. Для досягнення цієї мети були сформульовані та вирішені такі основні завдання:

- здійснити порівняльний аналіз та дослідити ефективність методів навчання карт Кохонена з метою виявлення аспектів, які ускладнюють їхнє практичне застосування;
- розробити методи та алгоритми, які вдосконалюють процес самоорганізації карт Кохонена;
- розробити метод встановлення відповідності елементів впорядкованих карт Кохонена кластерам даних; дослідити можливість візуалізації кластерної структури даних із використанням цього методу;
- дослідити застосовність розроблених методів та алгоритмів для аналізу структурованих та неструктурованих даних;
- розробити багатопотоковий алгоритм навчання карт Кохонена для організації ефективних обчислень на багатоядерних процесорах;
- розробити математичне та програмне забезпечення видобування даних із використанням карт Кохонена та запропонованих методів у вигляді програмної бібліотеки для побудови прикладних систем підтримки прийняття рішень.

Об'єктом дослідження є процеси моделювання даних на основі карт Кохонена.

Предметом дослідження є індуктивні методи моделювання даних на основі самоорганізовних карт Кохонена, візуалізація кластерної структури даних та ефективні алгоритми навчання в обчислювальних середовищах із багатоядерними процесорами.

Методи дослідження. У дисертаційній роботі використовувались загальнонаукові методи дослідження (спостереження, аналізу, порівняння) та конкретні, притаманні теорії самоорганізовних карт Кохонена та видобування даних. Зокрема, для перевірки коректності процесів самоорганізації встановлювались рівень топологічного впорядкування та усереднена похибка квантування (E.A. Uriarte, F.D. Martin). Адекватність отриманих моделей даних забезпечується використанням принципу зовнішнього доповнення, притаманного індуктивним методам моделювання (О.Г. Івахненко). З огляду на відсутність математичного апарату строгого доведення коректності функціонування самоорганізовних карт Кохонена, окрім одновимірної випадку (M. Cottrell, J.C. Fort, G. Pages), перевірка отриманих результатів ґрунтувалась на загальноприйнятій практиці у видобуванні даних – використання синтезованих та реальних еталонних наборів даних (A. Asuncion, D.J. Newman, R. Quinlan, K. Ulrich, R.A. Fisher, M. Marshall, Z.Q. Hong, J.Y. Yang, S. Aeberhard).

Наукова новизна одержаних результатів полягає в науковому обґрунтуванні та побудові вдосконалених і нових методів самоорганізації моделей даних на основі карт Кохонена. Отримано такі наукові результати:

- уперше розроблено метод класифікації на основі карт Кохонена шляхом застосування методу комітетів до отриманих моделей даних, що забезпечило ефективний спосіб визначення надійності результатів у задачах прийняття рішень, покладаючись на відсоток успішності моделюючих елементів карти;
- уперше розроблено метод “нейронної міграції” навчання карт Кохонена шляхом побудови відношення часткового порядку між елементами топологічних сусідств та відповідного переміщення елементів карти впродовж навчання, що забезпечило підвищення якості моделювання топології даних та успішності класифікації за тих самих умов та вибору параметрів, що й у методі Кохонена;
- уперше розроблено динамічно-інтервальну самоорганізовану карту шляхом використання інтервальних вагових векторів для моделюючих елементів, густини даних при навчанні та динамічної модифікації структури карти, що забезпечило ефективне усунення низки недоліків, притаманних картам Кохонена;
- удосконалено метод візуалізації кластерної структури даних високої розмірності на основі карт Кохонена шляхом застосування агломеративної кластеризації карти методом Unweighted Pair Group

Method with Arithmetic Mean (UPGMA) і присвоєння кольорових міток елементам карти, що забезпечило можливість узгодити візуалізацію даних із загальноприйнятим методом U-Matrix з одночасним визначенням елементів карти, які відповідають за моделювання даних з одного кластеру;

- подальшого розвитку дістав метод сегментації зображень на основі карт Кохонена шляхом використання повноколірної інформації та методу агломеративної кластеризації для визначення потрібних сегментів, що уможливило ефективне використання методу для відділення кисті руки від тла при вирішенні задачі побудови тренажера української жестової мови.

Практичне значення одержаних результатів. Практичне значення наукових результатів дисертаційного дослідження обумовлено підвищенням якості моделей даних, побудованих вдосконаленими та розробленими методами самоорганізації, що підтверджується експериментами на загальноприйнятих синтезованих наборах даних та успішністю впровадження систем прийняття рішень на їхній основі. Зокрема, практично цінними є результати:

- розроблено алгоритм декомпозиції і навчання карт Кохонена, який передбачає використання багатоядерної процесорної архітектури для розпаралелювання обчислень, що забезпечує ефективний спосіб масштабування обчислень, пов'язаних із самоорганізацією моделей даних;
- встановлено ефективність використання акторної моделі та технології MapReduce для організації паралельних обчислень шляхом порівняння ефективності програмної реалізації та зменшення часу виконання алгоритмів навчання, що уможливило вибір більш адекватної технології;
- розроблено математичне та програмне забезпечення комп'ютерної системи моделювання та аналізу даних у вигляді програмної бібліотеки для побудови прикладних систем підтримки прийняття рішень, яка інкорпорує розроблені методи та алгоритми моделювання і візуалізації даних;
- розроблено модуль проведення експериментів для перевірки ефективності навчальних алгоритмів та адекватності побудованих моделей даних, де для специфікації експериментів розроблено спеціалізовану мову опису на основі XML;
- розроблено модуль Веб-комунікації на основі ресурсної архітектури (англ. Resource Oriented Architecture, ROA) з метою збереження та обміну моделями даних через Інтернет.

Основні результати, отримані при виконанні дисертаційної роботи, використані для побудови комп'ютерної системи підтримки прийняття рішень у медичний діагностиці кардіозахворювань, яку впроваджено в кон-

сультативно-діагностичному центрі ТОВ “Десна” ЛТД (м. Тернопіль). Розроблений метод сегментації зображень використано з метою динамічного визначення зони зображення кистей рук на кадрах відеоряду при створенні комп’ютерної системи “Тренажер української жестової мови”, яку впроваджено для використання у Львівському навчальному інформаційному методичному центрі освіти, Львівському дитячому дистанційному навчально-консультаційному центрі дітей з особливими потребами на базі НВК “Школа-гімназія “Сихівська” та Львівській спеціальній загальноосвітній школі-інтернаті №101 Марії Покрови для глухих дітей.

Основні положення дисертації покладено в основу навчальних спеціалізованих курсів “Видобування даних” та “Основи теорії штучних нейронних мереж”, які викладаються на кафедрі дискретного аналізу та інтелектуальних систем Львівського національного університету імені Івана Франка. Розроблена комп’ютерна система моделювання та аналізу даних використовується студентами кафедри у курсових, дипломних та магістерських роботах.

Особистий внесок здобувача. Усі положення та результати дисертаційної роботи отримані автором самостійно. У роботах, опублікованих у співавторстві, здобувачеві належить: дослідження та застосування нейромереж прямого поширення [11, 14, 15, 18, 24, 26]; розроблення підходу до моделювання даних із використанням карт Кохонена та побудова класифікатора [10, 25]; дослідження теоретичних засад функціонування карт Кохонена [16]; запропонування та реалізація моделі синтезу нейромереж прямого поширення та карт Кохонена [9]; розроблення та обґрунтування теоретичних засад функціонування динамічно-інтервальної самоорганізованої карти, розроблення навчального алгоритму [12, 21]; обґрунтування теоретичних засад функціонування методу “нейронної міграції” [13]; схема дослідження впливу параметрів процесу самоорганізації на якість адаптації карт Кохонена, здійснення чисельних розрахунків і порівняльного аналізу, розроблення класифікатора даних [2]; розроблення методу визначення кластерних меж на адаптованій карті Кохонена, розроблення алгоритму візуалізації кластерної структури даних, дослідження підходів до візуалізації даних на карті Кохонена [5, 19, 20, 22]; розроблення підходу до аналізу структури даних на основі карт Кохонена [3]; розроблення методу кластеризації зображень за допомогою карт Кохонена, дослідження впливу вибору кольорових просторів при формуванні вхідних даних на якість сегментації [6, 4, 17, 23, 27]; розроблення індуктивних методів і алгоритмів аналізу структурованих даних та динамічної сегментації зображень на основі карт Кохонена [1].

Апробація результатів дисертації. Основні положення і результати дослідження доповідалися на 12 міжнародних та всеукраїнських наукових конференціях: International Conference on Inductive Modelling “ICIM-2002” (м. Львів, 2002) і “ICIM-2008” (м. Київ, 2008), всеукраїнській нау-

ковій конференції “Сучасні проблеми прикладної математики та інформатики” (м. Львів, 2003–2007), 6th International Conference on Information Systems Technology and its Applications “ISTA-2007” (м. Харків, 2007), VII международной конференции Интеллектуальный анализ информации “IAI-2007” (м. Київ, 2007), Experience of Designing and Application of CAD Systems in Microelectronics: IX-th International Conference “CADSM 2007” (м. Львів, м. Поляна, 2007), 3rd International United Information Systems Conference “UNISCON-2009” (м. Мельбурн, Астралія, 2009), IV міжвузівській науково-технічній конференції науково-педагогічних працівників “Проблеми та перспективи розвитку економіки і підприємства та комп’ютерних технологій в Україні” (м. Львів, 2009). У повному обсязі дисертаційна робота доповідалася та обговорювалася на наукових семінарах кафедри дискретного аналізу та інтелектуальних систем Львівського національного університету імені Івана Франка та кафедри інформаційних систем і мереж Національного університету “Львівська політехніка”.

Публікації. Результати дисертаційної роботи відображено у 27 наукових публікаціях. З них одна монографія, 10 статей у фахових наукових виданнях з технічних наук, 5 статей у фахових збірниках з фізикоматематичних наук, 11 публікацій матеріалів і тез доповідей міжнародних та всеукраїнських наукових конференцій.

Структура дисертації. Дисертаційна робота складається зі вступу, п’яти розділів, висновків, списку використаних джерел та додатку. Обсяг дисертації 171 сторінка (з них 126 сторінок основного тексту), містить 57 рисунків і 21 таблицю, які розташовані на 65 сторінках. Список використаних джерел складається з 171 найменування і займає 17 сторінок.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дослідження, визначено мету і завдання роботи, вказано методи дослідження, сформульовано наукову новизну та практичну цінність отриманих результатів, подано відомості про апробацію результатів та публікації, висвітлено особистий вклад здобувача в публікаціях із співавторами.

У **першому розділі** охарактеризовано сучасний стан теорії та інструментів аналізу даних. Виділено основні етапи виявлення знань із даних, задано контекст дисертаційного дослідження й означено використані головні поняття. Зокрема, під моделлю даних розуміємо алгебраїчну систему $AS = \langle X, \Omega_F, \Omega_P \rangle$ (А.І. Мальцев), де X – непорожня множина даних, Ω_F – множина операцій, визначених на множині X , Ω_P – множина предикатів, заданих для множини X . Множину X називають *носієм* системи AS , а її елементи – *елементами системи*. Пара $\langle X, \Omega_F \rangle$ – алгебра, яка визначає мову маніпулювання даними, а пара $\langle X, \Omega_P \rangle$ – модель або реляційна система, яка визначає мову опису моделі даних (Л.А. Калиниченко).

Показано зв'язок самоорганізовних карт Кохонена із загальною теорією самонавчання (Ф. Розенблат, М. Шлезінґер, В. Главач). Процес самонавчання визначено рекурентним співвідношенням $\tilde{K}_t = R(\tilde{X}, L(\tilde{X}, \tilde{K}_{t-1}))$, де t – поточний момент часу. Для карт Кохонена маємо: $\tilde{X}$ – вхідні дані, які моделюємо; $\tilde{K}$ – елементи найкращого наближення; R – алгоритм розпізнавання, який використовує стратегію пошуку елемента найкращого наближення; L – алгоритм навчання, який реалізує модифікацію вагових векторів $w \in W$ елементів карти. Таким чином, під *самоорганізацією моделі даних* розуміємо процес побудови чи модифікації моделі в результаті застосування алгоритмів самонавчання.

Подано результати системного аналізу сучасних напрямків розвитку теорії карт Кохонена (М. Cottrell, Н. Bauer, Т. Villman, J. Blackmore, В. Fritzke, М.С. Su, Н.Т. Chang, М. McInerney, А. Dhawan, М.С. Su, Н.Т. Chang, S.A. Harp, Т. Samad, S.J. Huang, С.С. Hung) та їхнього застосування для вирішення прикладних проблем (G. Deboeck, Т. Kohonen, С. Cinca, Е. Ainsworth, М. del Brio В., А. Varfis, С. Versino, J. Koh, М. Suk, S. Bhandarkar).

Зроблено висновок: застосування технологій аналізу даних на основі теорії карт Кохонена є актуальною задачею; існує необхідність розвитку методів та алгоритмів самоорганізації моделей даних та вдосконалення методів візуалізації даних високої розмірності з використанням карт Кохонена; існує потреба побудови алгоритмів навчання карт для виконання у багатопроцесорних обчислювальних середовищах для їхнього масштабування.

У **другому розділі** подано та проаналізовано властивості карт Кохонена, досліджено побудову методу класифікації на їхній основі та вибір оптимальних параметрів у методах навчання. Формально, самоорганізовною картою Кохонена називають нелінійне, впорядковане та гладке відображення $\Phi : X \rightarrow M$ елементів метричного простору високої розмірності на елементи регулярного масиву нижчої розмірності (переважно 1 та 2). Таке відображення будують відповідно до ітеративної процедури, яку називають методом навчання (Т. Kohonen, S. Haykin, Н. Ritter). Виділимо такі властивості відображення Φ : апроксимує вхідний простір, зберігає топологію та густину розподілу даних. Уважають, що в двовимірному випадку елементи $m_i \in M$ ($i = \overline{1, n}$, $|M| = n$) утворюють ґратку і розміщені у її вузлах. Конфігурація ґратки визначається взаємним розміщенням $m_i \in M$, які володіють цілочисловими координатами і утворюють евклідов простір. *Переможцем* для вхідного вектора $x \in X$ називають елемент $m(x) \in M$, для якого виконується умова $m(x) = \arg \min_i d(x, w_i)$, де w_i – ваговий вектор елемента $m_i \in M$ із того ж простору, що й вхідні дані, $d(x, w_i)$ – метрика вхідного простору X . Вектори w_i називають моделюючими (Т. Kohonen).

З метою підвищення практичності результату, дослідження ефективності навчання карт здійснено у контексті реальної задачі пошуку закономірностей в даних, які були зібрані впродовж 17 років при обстеженні пацієнтів на виявлення передумов виникнення певного кардіозахворювання у клініці Львівського державного медичного університету імені Данила Галицького. Досліджено та встановлено адекватність моделей даних, отриманих методом навчання Кохонена, методом стохастичної релаксації та методом нейронного газу (S. Osowski).

Для визначення точності моделей даних, побудованих у процесі самоорганізації карт Кохонена, використано рівень топологічного впорядкування та усереднена похибка квантування (Т. Kohonen, Е.А. Uriarte, F.D. Martin). Перевірку адекватності моделей здійснено на основі принципу зовнішнього доповнення (О.Г. Івахненко, Н. Madala), який реалізовувався шляхом розбиття усієї доступної множини даних X на навчальну та перевіральну підмножини із застосуванням критеріїв оцінки точності на перевіральній множині.

У ході дослідження розроблено класифікатор (див. алг. 1), який характеризується критерієм S – відсоток успішності класифікації даних, що поряд з іншими використано для визначення адекватності побудованих моделей. Складність запропонованого алгоритму $O(|M| \times |T|)$. Отримано приріст якості класифікації близько 10%: для відображення Φ , побудованого методом Кохонена без оптимальних значень параметрів, отримано успішність 83.61% та 81.92% (навчальна та перевіральна множини, відповідно), для оптимальних параметрів – 91.43% на перевіральній множині.

Алгоритм 1 Класифікатор.

Ініціалізація. Визначаємо пару змінних (a_i, b_i) для кожного елемента $m_i \in M$ для збереження кількості реагувань елемента на хворих та здорових пацієнтів. Присвоюємо $a_i \leftarrow 0, b_i \leftarrow 0, i = 1, |M|$. Вибраємо множину вхідних даних, для якої обчислюватиметься успішність класифікації, яку позначимо T . Множина елементів-переможців $K \leftarrow \emptyset$.

1. Вибраємо вектор $x \in T$ і вилучаємо його: $T \leftarrow T \setminus \{x\}$.
2. Для вектора x визначаємо переможця $m(x)$ і $K \leftarrow K \cup \{m(x)\}$.
3. Якщо вектору x відповідає ознака прийняття рішень зі значенням 1 (хворі пацієнти), то $a_{m(x)} \leftarrow a_{m(x)} + 1$, інакше – $b_{m(x)} \leftarrow b_{m(x)} + 1$.
4. Якщо $T \neq \emptyset$, то повертаємось на крок 1.
5. Обчислюємо успішність кожного елемента $m \in K$:

$$s_k = \frac{\xi_k}{a_k + b_k} * 100, k \in I(K), \text{ де } I(K) - \text{множина індексів елементів-переможців,}$$

$$\xi_k = \begin{cases} a_k, & a_k \geq b_k; \\ b_k, & a_k < b_k. \end{cases}$$

6. Обчислюємо загальну успішність класифікації: $S = \frac{1}{|K|} \sum_{k \in I(K)} s_k$, де $|K|$ – кількість елементів-переможців для векторів множини T .
-

У третьому розділі подано основні результати дослідження щодо вдосконалення процесів самоорганізації моделей даних. Якість моделей

даних, отриманих у результаті самоорганізації карт Кохонена, залежить від ініціалізації вагових векторів та параметрів навчальних алгоритмів. Цей висновок підтверджено результатами цієї дисертаційної роботи та низкою праць, присвячених дослідженню питань ініціалізації карт Кохонена та побудові альтернативних підходів до навчання (Z.-P. Lo, B. Bavarian, S.A. Harp, T. Samad, M.C. Su, H.T. Chang, S.J. Huang, C.C. Hung, J. Blackmore, H. Bauer, T. Villman та інші).

Головною метою методу “нейронної міграції” (НМ) є підвищення топологічного впорядкування карти шляхом переформування елементів ґратки впродовж етапу організації методу Кохонена. Адаптивне перегрупування елементів карти у методі НМ проводиться відповідно до близькості їхніх вагових векторів. Процес перегрупування елементів названо нейронною міграцією. Структурно метод НМ складається з двох процесів: нейронної міграції та адаптації, де адаптація відбувається відповідно до методу Кохонена.

Топологічним околом елемента-переможця m_c для вхідного вектора $x \in X$ є множина $O_\varepsilon(m_c) = \{m \in M \setminus \{m_c\} \mid h_{m,m_c} > \varepsilon\}$, де константа $\varepsilon \geq 0$, h_{m,m_c} – значення функції топологічного сусідства, M – вихідний простір. Топологічний окіл використовується для обмеження кількості елементів, що беруть участь у міграції. Це необхідно для уникнення міграції елементів між топологічними сусідствами різних елементів-переможців. *Відношенням міграції* на $O_\varepsilon(m_c)$ називають множину $R_{m_c} = \{(m_k, m_l) \mid d(w_k, w_c) \leq d(w_l, w_c) \wedge d_{k,c} \geq d_{l,c}\}$, де w_k, w_l, w_c – вагові вектори елементів $m_k \in O_\varepsilon(m_c)$, $m_l \in O_\varepsilon(m_c)$ та m_c , відповідно, $d(\cdot, \cdot)$ – метрика вхідного простору X , $d_{\cdot, \cdot}$ – метрика вихідного простору M . Сформульовано і доведено твердження, що R_{m_c} є відношенням часткового порядку. На основі цього твердження розроблено алгоритми топологічного сортування множини $(O_\varepsilon(m_c), R_{m_c})$ та міграції її елементів, які є основою алгоритму НМ (див. алг. 2).

Адекватність методу НМ перевірено на низці еталонних наборів даних (A. Asuncion, D.J. Newman). Отримані результати вказують на 6% приросту успішності класифікації з одночасним зниженням топологічної помилки. Обчислювальна складність алгоритму міграції в найгіршому випадку є $O(|O_\varepsilon(m_c)|^2)$, а у найкращому – $O(|O_\varepsilon(m_c)| \log |O_\varepsilon(m_c)|)$.

Практичне використання карт Кохонена вказує на такі недоліки: статична, наперед визначена структура нейронної ґратки призводить до неадекватної апроксимації даних за умови невеликого вибору кількості елементів та топології; відображення Φ забезпечує “точкову” апроксимацію даних, що ускладнює виявлення хибної класифікації; неможливість донавчання карт Кохонена – у практичних задачах виникає потреба уточнення моделі даних, що у випадку карт Кохонена призводить до повної перебудови відображення Φ . Для вирішення цих питань розроблено динамічно-інтервальну самоорганізовану карту (ДІСК), в основу якої покладено теорію карт Кохонена та засади інтервального аналізу. Елементи ДІСК ха-

рактизуються трійкою величин (W_i, S_i, w_i) : W_i – інтервальний ваговий вектор, що визначає гіперкубічну область вхідного простору, за моделювання даних якої відповідає елемент; $S_i = (S_i^1, \dots, S_i^m)$ – вектор, компонентами якого є множини, відповідні інтервальним компонентам вектора $W_i = ([\min(S_i^1), \max(S_i^1)], \dots, [\min(S_i^m), \max(S_i^m)])$, що є одним із трактувань числового інтервалу (Е. Hansen, G.W. Walster); $w_i = (\overline{S_i^1}, \dots, \overline{S_i^m})$ – точковий ваговий вектор елемента карти, компонентами якого є середні значення елементів вектора S_i .

Алгоритм 2 Нейронна міграція: алгоритм навчання карт Кохонена.

Ініціалізація. За початкові значення вагових векторів $w_k(0)$ ($k = \overline{1, |M|}$) вибираємо вектори, компонентам яких присвоєно випадкові числа з проміжку $[0, 1]$. При цьому має виконуватись умова $w_k(0) \neq w_l(0)$ для $k, l = \overline{1, |M|}$, $k \neq l$. Вибираємо $\varepsilon > 0$ для утворення топологічного околу елемента-переможця. Покладаємо $n \leftarrow 1$ – номер поточної ітерації. Покладаємо κ рівне половині кількості ітерацій фази впорядкування.

1. Вибираємо вхідний вектор $x \in X$ випадковим чином.
 2. Для вектора x визначаємо переможця m_c .
 3. Якщо $n \leq \kappa$, то утворюємо топологічний окіл $O_\varepsilon(m_c)$, до якого застосовуємо алгоритм міграції.
 4. Модифікуємо вагові вектори елементів карти відповідно до співвідношень методу Кохонена.
 5. Покладаємо $n \leftarrow n + 1$.
 6. Перевіряємо умову зупинки. Якщо умова зупинки не виконується, то перехід на крок 1.
-

Формування карти відбувається динамічно впродовж навчання, починаючи з одного елемента. З огляду на громіздкість алгоритму, далі висвітлюємо лише головні етапи методу навчання ДІСК.

1. *Змагання.* При поданні вхідного вектора $x = (x^1, \dots, x^m) \in X$ для кожного елемента обчислюється деяка класифікуюча функція. Елемент із найбільшим значенням – переможець.
2. *Адаптація.* Етап адаптації спрямований на модифікацію переможця i для підвищення його інформаційної насиченості, що обчислюється деякою функцією $f(W_i, S_i)$. Під модифікацією елемента розуміємо переобчислення трійки його величин відповідно до співвідношень: $S_i^j \leftarrow S_i^j \cup \{x^j\}$, $W_i^j \leftarrow [\min(S_i^j), \max(S_i^j)]$, $w_i^j \leftarrow \frac{\sum_{s \in S_i^j} s}{|S_i^j|}$ ($j = \overline{1, m}$). Якщо у результаті модифікації (W_i, S_i, w_i) інформаційна насиченість зменшилась, то елемент розщеплюють на два, що веде до нарощення нових елементів карти. Вибір конкретної функції $f(W_i, S_i)$ визначає динаміку цього процесу.
3. *Взаємодія.* Під взаємодією елементів розуміємо їхнє об'єднання з метою утворення одного елемента з вищим рівнем інформаційної насиченості.

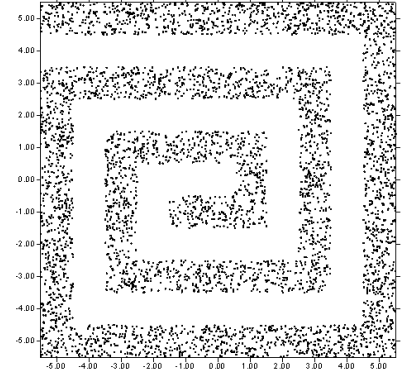
Використання інтервальних вагових векторів W_i уможливорює контроль за хибною класифікацією. Навчальний метод визначає динаміку зміни

карти з підтримкою її донавчання. Приклад застосування ДІСК подано на рис. 1. Усереднена похибка квантування склала 0.21602 для ДІСК та 0.30606 для карти Кохонена, навченої на тому ж наборі даних без підбору оптимальних параметрів².

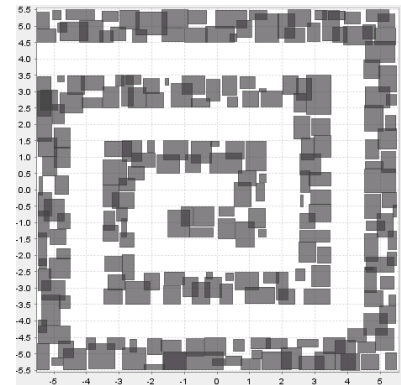
У четвертому розділі досліджено підходи до візуального аналізу даних з використанням карт Кохонена, які у випадку двовимірної ґратки елементів уможливають зображення кластерної структури даних високої розмірності. Одним із поширених методів візуалізації є U-Matrix (A. Ultsch, H.P. Siemon). У результаті його застосування формується двовимірне зображення кластерної структури, але не вказується, які елементи карти відповідають за відображення кластерів.

З метою інтерпретації навчених карт Кохонена та візуалізації кластерних структур даних розроблено метод, який визначає відповідність між елементами карти і кластерами даних, а також візуалізує структуру даних у відповідності з методом U-Matrix. Центральною ідеєю методу є використання агломеративної кластеризації (UPGMA) елементів карти з використанням метрики вхідного простору. Результатом агломеративної кластеризації є дендрограма, до якої застосовується переріз на висоті l з метою утворення множини кластерів $C(l)$. Елементами кластерів є елементи карти Кохонена. Кожному кластеру $C_i \in C(l)$ присвоюють мітку (H_i, S, L) з кольорового простору HSL. Значення S та L покладаються деякими сталими для усіх міток, а значення H_i обчислюють відповідно до співвідношень $H_i \leftarrow i \times \Delta H_i$ та $\Delta H_i \leftarrow \frac{\max(H) - \min(H)}{|C(l)|}$ ($i = \overline{1, |C(l)|}$). Після присвоєння кольорових міток елементи двовимірної карти можна зобразити на площині. Таким чином, запропонований метод є кластеризацією карти Кохонена з підтримкою візуалізації завдяки присвоєнню кольорових міток. Алгоритм 3 реалізує описаний метод. Приклад результату застосування алгоритму на еталонному наборі даних Lung Cancer (Z.Q. Hong, J.Y. Yang) зображено на рис. 2 у порівнянні з результатом методу U-Matrix.

Розроблений метод кластеризації карт Кохонена з метою візуалізації структур даних успішно використано для сегментації зображень у задачі побудови тренажера української мови жестів. Метою сегментації було



(a) Вхідні дані



(b) Модель даних

Рис. 1. Приклад застосування ДІСК.

²Архітектура ДІСК не вимагає налаштування параметрів.

відділення фону зображення від сегменту, який репрезентує руку. Причому, цей процес мав відбуватись динамічно для кожного кадру відеопотоку впродовж взаємодії користувача з тренажером.

Алгоритм 3 Маркування карти Кохонена.

Ініціалізація. W – множина вагових векторів навченої карти Кохонена; U – карта висот, отримана для цієї ж карти Кохонена методом U-Matrix.

1. Застосовуємо алгоритм UPGMA до множини W , результатом чого є дендрограма D .
 2. Вибираємо рівень перерізу l дендрограми D .
 3. Здійснюємо переріз D на рівні l , результатом чого є множина кластерів $C(l)$.
 4. Утворюємо кольорові мітки (H_i, S, L) для кластерів $C_i \in C(l)$ ($i = \overline{1, |C(l)|}$) відповідно до формул методу.
 5. Присвоюємо мітки елементам карти m_j ($j = \overline{1, |W|}$) відповідно до їхньої належності кластерам $C_i \in C(l)$ ($i = \overline{1, |C(l)|}$).
 6. Зображаємо 2D карту, де кожен елемент має колір присвоєної йому мітки.
 7. Порівнюємо зображену промарковану карту з картою висот U ; якщо результат адекватний, то кінець – елементи з однаковими мітками відповідають одному кластеру; інакше – перехід на крок 2.
-

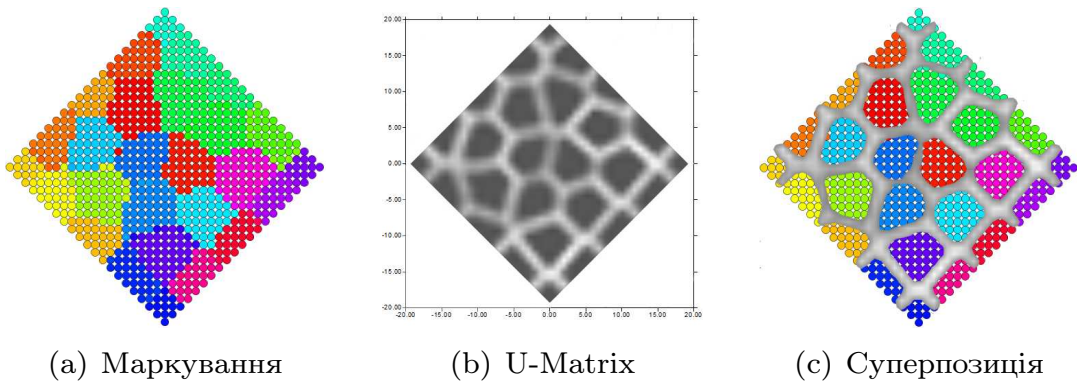


Рис. 2. Результат маркування.

Одним із головних аспектів здійсненого дослідження було формування та підготовка даних для навчання карти Кохонена, що вимагало аналізу впливу різних кольорових просторів (R.W.G. Hunt, G. Hoffmann) на якість кластеризації зображень. Особливістю запропонованого методу сегментації зображень є використання простору CIELab та зменшення навчальної вибірки з метою прискорення навчання карти Кохонена. Вищерозглянутий алгоритм кластеризації карти Кохонена використано з метою отримання сегментів зображення для виділення лише двох – сегмента фону та сегмента руки. Важливо зауважити, що навчання карти Кохонена використовувалось лише на першому кадрі – на решті використовувалась уже сформована карта. Укрупнена схема розробленого алгоритму сегментації зображена на рис. 3. Приклад застосування сегментації до послідовності кадрів відеоряду подано на рис. 4.

У розділі п'ять подано деталі розробленого програмного забезпечення, яке інкорпорує низку модулів, що забезпечують гнучке та ефективне

розроблення прикладного програмного забезпечення підтримки прийняття рішень, а також з метою створення спеціалізованих програмних бібліотек аналізу даних. На рис. 5 зображено компонентну діаграму розробленого програмного забезпечення із виділеними допоміжними модулями ресурсної взаємодії через Інтернет (R. Fielding) і об'єктно-орієнтованої взаємодії із системами керування реляційними базами даних (G. King).

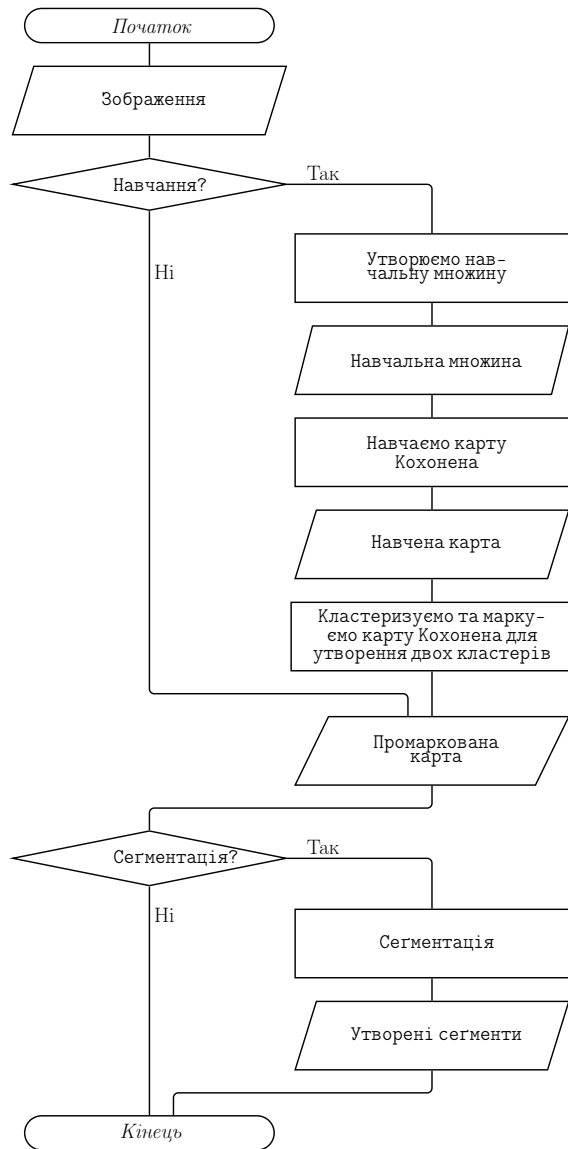


Рис. 3. Блок-схема алгоритму сегментації зображення.

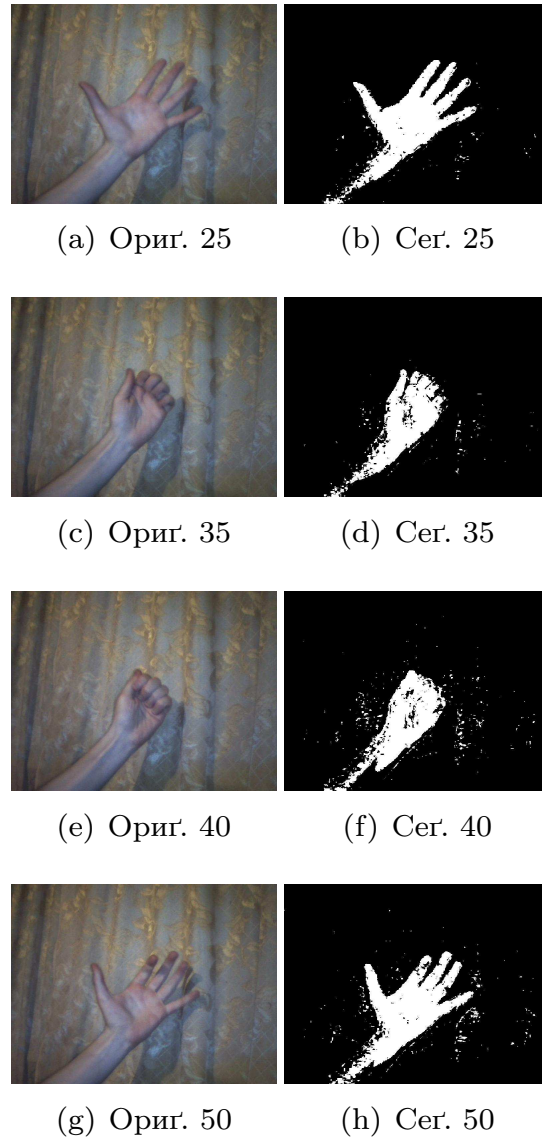


Рис. 4. Оригінальні та сегментовані кадри 25, 35, 40 і 50.

З огляду на тенденцію розвитку мікропроцесорної архітектури у бік багатоядерних процесорів, актуальним є питання масштабованості розробленого математичного та програмного забезпечення. Для підтримки масштабованості процесу моделювання даних розроблено незалежний від топології ґратки багатопотоковий алгоритм навчання карт Кохонена. Ефективність реалізації розробленого алгоритму досліджено з використанням двох

сучасних парадигм паралельних обчислень – акторної моделі (М. Oderski) та технології MapReduce (J. Dean, S. Ghemawat). Порівняння часу виконання навчального процесу на еталонному наборі Lung Cancer для різної кількості груп елементів подано на рис. 6. Результат отримано на обчислювальній техніці з процесором Intel Xeon E5504, який має чотири ядра з частотою 2GHz кожне.

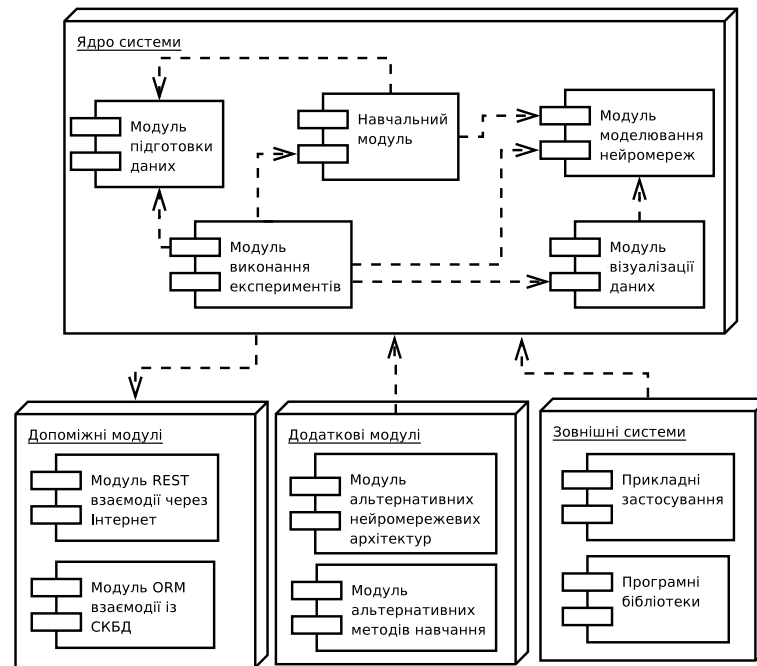


Рис. 5. Компонентна діаграма розробленого програмного забезпечення.

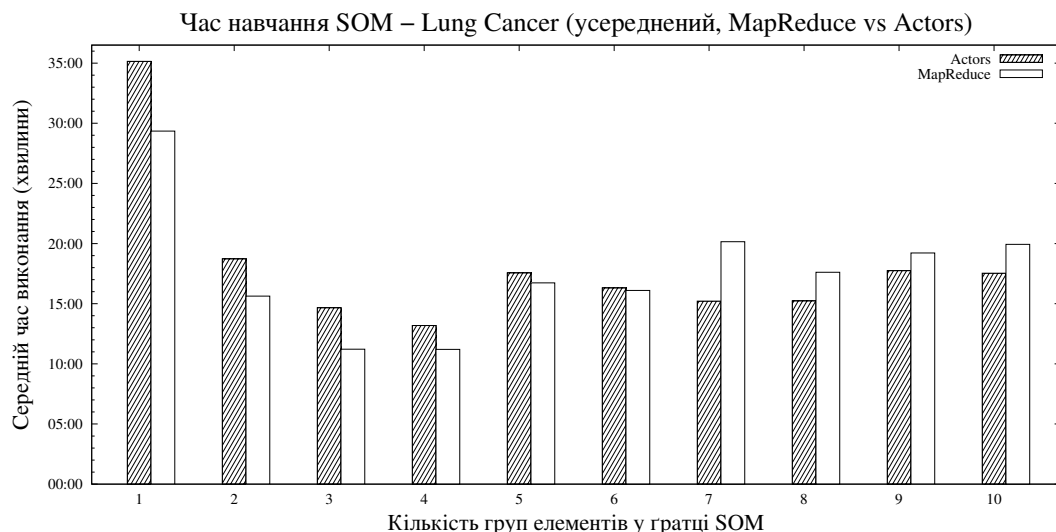


Рис. 6. Порівняння ефективності обчислень у багатоядерному середовищі.

Усі висвітлені в дисертації результати чисельних експериментів отримано з використанням спеціально розробленого модуля виконання експериментів. Сценарій експериментів є XML документом, у якому можна визначати джерело та формат даних, метод та параметри навчання тощо.

Ядро розробленого програмного забезпечення складається із 8761 рядка коду. Програмний каркас утворено з 27 базових інтерфейсів. Середнє значення та стандартна девіація цикломатичного числа склала 1.683, максимальне значення – 7, що не перевищує рекомендованих норм (Thomas J. McCabe). При розробленні програмного забезпечення за мету поставлено використання виключно відкрите та вільне програмне забезпечення. Уся робота виконана на операційній системі Ubuntu Linux із використанням мов програмування Java 1.6 та Scala 2.7, середовища розробки Eclipse і низки допоміжних бібліотек з відкритим кодом.

ОСНОВНІ РЕЗУЛЬТАТИ ТА ВИСНОВКИ

У дисертаційній роботі вирішено актуальну науково-технічну задачу розвитку індуктивних методів аналізу даних на основі самоорганізовних карт Кохонена і розроблення ефективного математичного та програмного забезпечення комп'ютерної системи аналізу даних. Основні наукові та практичні результати полягають у наступному.

1. Здійснено порівняльний аналіз та досліджено ефективність методів навчання карт Кохонена, завдяки чому виявлено причини складності їхнього практичного застосування і розроблено підходи оцінювання параметрів навчальних алгоритмів, що забезпечило приріст якості класифікації у задачі діагностування кардіозахворювань з 81.92% до 91.43% на перевіряльній вибірці.
2. Розроблено метод класифікації на основі карт Кохонена шляхом застосування методу комітетів до отриманих моделей даних, що забезпечило ефективний спосіб визначення надійності результатів у задачах прийняття рішень, покладаючись на відсоток успішності моделюючих елементів карти.
3. Розроблено метод “нейронної міграції” навчання карт Кохонена, який ґрунтується на ідеї підвищення топологічного впорядкування карти, та розроблено динамічно-інтервальну самоорганізовну карту шляхом використання інтервальних вагових векторів для моделюючих елементів, густини даних при навчанні та динамічної модифікація структури карти – ці методи забезпечили подолання низки недоліків, притаманних картам Кохонена, зокрема, проблеми ініціалізації, що забезпечило істотний приріст якості класифікації.
4. Удосконалено метод візуалізації кластерної структури даних високої розмірності на основі карт Кохонена шляхом застосування агломеративної кластеризації карти методом UPGMA та присвоєння кольорових міток елементам карти, що забезпечило можливість узгодити візуалізацію даних із загальноприйнятим методом U-Matrix з одночасним визначенням елементів карти, що відповідають за моделювання даних з одного кластеру.

5. Розроблено метод сегментації зображень на основі карт Кохонена з використанням повноколірної інформації та методу агломеративної кластеризації для визначення потрібних сегментів, що уможливило використання методу для відділення кисті руки від тла при вирішенні задачі побудови тренажера української жестової мови. Реалізовано ефективний алгоритм методу, який використовує кешування та багатопотокове опрацювання кадрів відеоряду.
6. Розроблено алгоритм багатопотокового навчання карт Кохонена, який на системі з чотириядерним процесором забезпечує 180% приріст швидкодії порівняно з послідовним навчальним алгоритмом, що дало змогу суттєво скоротити час побудови моделей даних.
7. Розроблено математичне та програмне забезпечення видобування даних у вигляді програмної бібліотеки, яка містить усі методи і алгоритми, розроблені в дисертації, а також низку важливих модулів, що забезпечують Веб-комунікацію, планування та проведення експериментів. У розробленні використано виключно технології з відкритим кодом на основі мов програмування Java та Scala.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. *Годич О.* Українська жестова мова: комп'ютерно-лінгвістичний аспект: Монографія / О. Годич, М. Давидов, Ю. Нікольський, В. Пасічник, Ю. Щербина. — Львів: “Літературна агенція “Піраміда”, 2009. — 253 с.
2. *Годыч О.* Исследование эффективности алгоритмов обучения нейросетей Кохонена / О. Годыч, В. Пасичник, Ю. Никольский, Ю. Щербина // *Управляющие системы и машины*. — 2006. — № 2. — С. 63–80.
3. *Годич О.* Аналіз структури медичних даних із застосуванням мереж Кохонена / О. Годич, Ю. Нікольський, В. Пасічник, Ю. Щербина // *International Journal of Computing*. — 2007. — Т. 6, № 3. — С. 124–136.
4. *Hodych O.* SOM-based dynamic image segmentation for sign language training simulator / O. Hodych, K. Hushchyn, I. Nikolski, V. Pasichnyk, Y. Shcherbyna // *Information Systems: Modeling, Development, and Integration* / Ed. by W. van der Aalst, J. Mylopoulos, N. M. Sadeh, M. J. Shaw, C. Szyperski. — Berlin: Springer Berlin Heidelberg, 2009. — Vol. 20 of *Lecture Notes in Business Information Processing*. — P. 29–40.
5. *Hodych O.* Determining cluster boundaries within Self-Organizing Maps / O. Hodych, I. Nikolski, V. Pasichnyk, Y. Shcherbyna // *Вісник Національного технічного ун-ту “Харківський політехнічний інститут”*. — 2007. — № 5. — С. 97–109.
6. *Годыч О.* Динамическая сегментация изображений для учебного симулятора языка жестов / О. Годыч, К. Гушин, Ю. Никольский, В. Па-

- сичник, Ю. Щербина // *Управляющие системы и машины*. — 2009. — № 1. — С. 79–85.
7. Годич О. Навчання SOM методом нейронної міграції / О. Годич // *Вісник Національного ун-ту “Львівська політехніка”, Інформаційні системи та мережі*. — 2004. — № 519. — С. 55–72.
 8. Годич О. Кластеризація даних нейромережею ADD / О. Годич // *Вісник Національного ун-ту “Львівська політехніка”, Інформаційні системи та мережі*. — 2005. — № 549. — С. 54–68.
 9. Hodych O. Synthesis of self-organizing map and feedforward neural network for better forecasting / O. Hodych, Y. Shcherbyna, M. Zylan // *International Journal of Computing*. — 2004. — Vol. 3, no. 3. — P. 68–75.
 10. Годич О. Застосування штучної нейронної мережі типу SOM для розв’язування задачі діагностування / О. Годич, Ю. Нікольський, Ю. Щербина // *Вісник Національного ун-ту “Львівська політехніка”, Інформаційні системи та мережі*. — 2002. — № 464. — С. 31–43.
 11. Щербина Ю. Методи навчання штучної нейронної мережі / Ю. Щербина, О. Годич // *Вісник Національного ун-ту “Львівська політехніка”, Інформаційні системи та мережі*. — 2001. — № 438. — С. 160–170.
 12. Годич О. Динамічна нейромережа ADD / О. Годич, Ю. Щербина // *Вісник Львівського національного ун-ту ім. І.Франка, серія “Прикладна мат. та інформ.”*. — 2005. — № 10. — С. 161–183.
 13. Годич О. Метод нейронної міграції для навчання нейромереж типу SOM / О. Годич, Ю. М. Щербина // *Вісник Львівського національного ун-ту ім. І.Франка, серія “Прикладна мат. та інформ.”*. — 2006. — № 11. — С. 203–209.
 14. Годич О. Застосування штучних нейронних мереж для прогнозування курсу акцій / О. Годич, Б. Голуб, Ю. Щербина // *Вісник Львівського національного ун-ту ім. І.Франка, серія “Прикладна мат. та інформ.”*. — 2002. — № 4. — С. 152–168.
 15. Годич О. Застосування штучних нейронних мереж до розв’язування задач про найменші квадрати / О. Годич, Ю. Щербина // *Вісник Львівського національного ун-ту ім. І.Франка, серія “Прикладна мат. та інформ.”*. — 2003. — № 6. — С. 182–190.
 16. Годич О. Самоорганізація нейромереж та класифікація даних / О. Годич, Ю. Щербина // *Вісник Львівського національного ун-ту ім. І.Франка, серія “Прикладна мат. та інформ.”*. — 2003. — № 7. — С. 234–247.
 17. Hodych O. The dynamic image segmentation for sign language training simulator / O. Hodych, K. Hushchyn, I. Nikolski, V. Pasichnyk, Y. Shcherbyna // 2nd International Conference on Inductive Modelling, Kyiv, Ukraine, September 15 – 19, 2008: Proceedings. — Kyiv, 2008. — P. 202–206.
 18. Нікольський Ю. Застосування штучних нейронних мереж для розв’язування задач прогнозування часових послідовностей / Ю. Ніколь-

- ський, О. Годич, Ю. Щербина // Праці Міжнародної конференції з індуктивного моделювання (ICIM-2002). — Львів, 2002. — С. 144–149.
19. *Hodych O.* High-dimensional data structure analysis using Self-Organising Maps / O. Hodych, I. Nikolski, V. Pasichnyk, Y. Shcherbyna // Proceedings of 9th International Conference on CAD Systems in Microelectronics (CADSM 2007). — Lviv-Polyana, 2007. — P. 218–221.
 20. *Годыч О.* Применение нейросетей Кохонена для выявления и визуализации скрытых структур в данных высокой размерности / О. Годыч, Ю. Никольский, В. Пасичник, Ю. Щербина // Тез. докл. VII междунар. конф. “Интеллектуальный анализ информации-2007” (IAI-2007). — Киев, 2007. — С. 45–55.
 21. *Годич О.* Динамічна нейромережа ADD / О. Годич, Ю. Щербина // Тез. доп. XII Всеукр. наук. конф. “Сучасні проблеми прикл. матем. та інформ.”. — Львів, 2005. — С. 65–66.
 22. *Годич О.* Визначення кластерних меж у SOM / О. Годич, Ю. Щербина // Тез. доп. XIII Всеукр. наук. конф. “Сучасні проблеми прикл. матем. та інформ.”. — Львів, 2006. — С. 42.
 23. *Годич О.* Динамічна сегментація зображень на основі SOM / О. Годич, К. Гуцин, Ю. Щербина // Тез. доп. XIV Всеукр. наук. конф. “Сучасні проблеми прикл. матем. та інформ.”. — Львів, 2007. — С. 49–50.
 24. *Годич О.* Застосування штучної нейромережі до розв’язування нелінійних задач про найменші квадрати / О. Годич, Ю. Щербина // Тез. доп. IX Всеукр. наук. конф. “Сучасні проблеми прикл. матем. та інформ.”. — Львів, 2002. — С. 32–33.
 25. *Годич О.* Модифікація геометричної структури нейронних мереж типу SOM / О. Годич, Ю. Щербина // Тез. доп. X Всеукр. наук. конф. “Сучасні проблеми прикл. матем. та інформ.”. — Львів, 2003. — С. 45–46.
 26. *Годич О.* Застосування нейромереж в комбінаторних оптимізаційних задачах / О. Годич, Ю. Щербина // Тез. доп. XI Всеукр. наук. конф. “Сучасні проблеми прикл. матем. та інформ.”. — Львів, 2004. — С. 46–47.
 27. *Годич О.* Динамічна сегментація зображень на основі самоорганізаційних карт Кохонена / О. Годич, Ю. Щербина // Збірник матеріалів IV міжвузівської науково-технічної конференції науково-педагогічних працівників “Проблеми та перспективи розвитку економіки і підприємства та комп’ютерних технологій в Україні”. — Львів, 2009. — С. 42–43.

АНОТАЦІЇ

Годич О. В. Індуктивні методи та алгоритми самоорганізації моделей даних на основі карт Кохонена. — Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 01.05.03 — математичне та програмне забезпечення

обчислювальних машин і систем. — Національний університет “Львівська політехніка”, Львів, 2010.

Дисертацію присвячено процесам моделювання даних на основі карт Кохонена. На основі розробленої методики визначення оптимальних значень параметрів навчальних методів досягнуто приріст успішності класифікації з 81.92% до 91.43%. Розроблено метод “нейронної міграції” навчання, що ґрунтується на ідеї підвищення топологічного впорядкування карт, та розроблено динамічно-інтервальну самоорганізовану карту, які забезпечили істотний приріст успішності класифікації. Удосконалено метод візуалізації кластерної структури даних, результати роботи якого узгоджуються з методом U-Matrix. На його основі розроблено алгоритм сегментації зображень, який використано при побудові тренажера української жестової мови. Розроблено математичне та програмне забезпечення для інтелектуального аналізу даних, яке інкорпорує багатопотоковий алгоритм навчання карт, що на системі з чотириядерним процесором забезпечує збільшення швидкодії у 2.8 раза, порівняно з послідовним алгоритмом.

Ключові слова: карти Кохонена, інтелектуальний аналіз даних, візуалізація даних, індуктивне моделювання, паралельні обчислення.

Годыч О. В. Индуктивные методы и алгоритмы самоорганизации моделей данных на основе карт Кохонена. — Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 01.05.03 — математическое и программное обеспечение вычислительных машин и систем. — Национальный университет “Львовская политехника”, Львов, 2010.

Диссертация посвящена процессам моделирования данных с использованием карт Кохонена. На основании разработанной методики определения оптимальных значений параметров обучающих методов получен прирост качества классификации с 81.92% до 91.43%. Разработан метод “нейронной миграции” обучения, который основан на идее повышения топологического упорядочения карт, а также динамично-интервальная самоорганизуемая карта, которые обеспечили существенный прирост качества классификации. Усовершенствован метод визуализации кластерной структуры данных, результаты работы которого согласуются с методом U-Matrix. На его основании разработан алгоритм сегментации изображений, который был использован при построении тренажера украинского жестового языка. Разработано математическое и программное обеспечение интеллектуального анализа данных, которое включает многопоточный алгоритм обучения, который на системе с 4-х ядерным процессором обеспечивает увеличение производительности в 2.8 раза, по сравнению с последовательным алгоритмом.

Ключевые слова: карты Кохонена, интеллектуальный анализ данных, визуализация данных, индуктивное моделирование, параллельные вычисления.

Hodych O. Inductive methods and algorithms for self-organising data modeling based on Kohonen’s maps. — Manuscript.

Candidate’s thesis on Technical Sciences, speciality 01.05.03 — mathematical and software support of computational machines and systems. — Lviv Polytechnic National University, Lviv, 2010.

The goal of this thesis is to advance the state of inductive data modeling methods and algorithms based on Kohonen’s self-organising maps (SOM) with an emphasis on exploratory data analysis, descriptive data models and data visualisation. All developed improvements and new methods have been evaluated on a number of datasets from UCI Machine Learning Repository. At the same time an important part of the research is related to the practical applications of the developed methods and algorithms, which cover cases of modeling both structured (medical applications) and unstructured (image segmentation) data, including visualisation. In light of the recent trends in the CPU industry moving towards multi-core architectures, the developed during the research software incorporates modifications of SOM training algorithms optimised for symmetric multiprocessing.

The first chapter of the thesis introduces all important definitions and provides an interpretation of SOM as an algebraic system provided with a self-organising learning process. It also contains a detailed overview of recent theoretical and practical advancements in SOM. The second chapter provides an in-depth review of the main SOM features and the results of the developed methodology for selecting optimal parameter values for a number of training algorithms in application to analysis of the real-life medical data. As the result the classification success has been increased from 81.92% to 91.43%. The details of a new training method “Neural Migration” and a new dynamic interval self-organising map are provided in chapter three. Their application to the test datasets yields significant improvements over alternatives. The improved algorithm for data visualisation based on SOM and a derived from it image segmentation method is discussed in the fourth chapter. The obtained results correlate with the results of U-Matrix, but in addition it determines groups of map elements responsible for modeling of individual data clusters. The last chapter provides a detailed overview of the developed data mining software and the proposed multi-threaded training algorithm, which yields on average up to 2.8 times performance improvement over the sequential version when executing on Intel Xeon E5504 CPU. An important aspect of the developed software is the use of strictly free and open source libraries, programming languages (Java 1.6 and Scala 2.7) and an environment (Eclipse IDE under Ubuntu OS). The Scala programming language was used for its functional features providing a flexible way for implementing actor-based parallel computations while remaining fully compatible with the Java code-base of the developed software.

Key words: Self-Organising Maps, data mining, data visualisation, inductive modelling, parallel computing.

Підписано до друку 19.08.2010. Формат 60 × 84/16. Папір друк. Офсет.
друк. Фіз. друк. арк. 1,5. Умовн. друк. арк. 1,4.
Тираж 100 пр. Зам. 89.

Товариство з обмеженою відповідальністю “Літературна агенція “Пірамі-
да”, 79024, м. Львів, вул. Промислова, 45.