УДК 004.853+004.855.5

О.В. Годич, Т.П. Мазепа
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ВІЗУАЛІЗАЦІЯ ДАНИХ, КЛАСТЕРИЗОВАНИХ ДИНАМІЧНО-ІНТЕРВАЛЬНОЮ САМООРГАНІЗОВНОЮ КАРТОЮ

Вирішується завдання візуалізації кластерної структури даних високої розмірності на основі моделі даних, отриманої динамічно-інтервальною самоорганізовною картою (ДІСК). За розробленим методом візуалізації використовують карту Кохонена з метою проектування елементів ДІСК на двовимірну ґратку, у поєднанні із алгоритмом U-Matrix для візуалізації кластерів даних.

Ключові слова: візуалізація даних, аналіз даних, самоорганізовні карти.

In this article we present an algorithm for visualising the clustering structure of the data model captured by dynamic interval self-organising map (DISOM). The developed visualisation algorithm employs the Self-Organising Map for placing DISOM elements on the 2D lattice in conjunction with U-Matrix algorithm for visualization of data clusters.

Keywords: data visualisation, data analysis, self-organising maps.

Вступ. Загальна постановка проблеми

Візуалізація кластерної структури даних високої розмірності є однією із центральних тем досліджень аналізу даних та інженерії знань [1, 2]. Самоорганізовні карти Кохонена є ефективною технологією кластеризації даних, які уможливлюють їхню візуалізацію [3–5]. Водночас під час практичного використання карт Кохонена виявлено на низку недоліків, які було досліджено у [6–14]:

1. Статична, наперед визначена структура нейронної ґратки, що призводить до неадекватної апроксимації даних за умови невдалого вибору кількості елементів і топології сусідства;
2. Відображення Φ, яке реалізує навчена карта Кохонена, забезпечує “точкову” апроксимацію даних, а тому завжди знайдеться елемент якнайкращого наближення, що ускладнює виявлення хибної класифікації;
3. Неможливість донавчання карт Кохонена; у практичних задачах виникає потреба уточнення моделі даних, що у випадку карт Кохонена призводить до повної перебудови відображення Φ.

Динамічно-інтервальну самоорганізовну карту (ДІСК) було розроблено з метою подолання цих недоліків. В основу ДІСК покладено ідеї процесу самоорганізації карт Кохонена та засади інтервального аналізу. Використання інтервальних вагових векторів в елементах ДІСК дає змогу побудувати гіперкубічні області простору даних для їхнього моделювання. Метод побудови інтервальних векторів використовує гіпотезу λ−компактності [15], що вирізняє технологію ДІСК серед інших досліджень. Детальне обговорення функціонування технології ДІСК і порівняння її ефективності із картами Кохонена висвітленно [6, 16–18].

Виділення проблеми. Формулювання мети

Одним із недоліків ДІСК порівняно з картами Кохонена є відсутність структурованої двовимірної ґратки її елементів. У випадку карт Кохонена це забезпечує можливість візуалізації кластерної структури модельованих даних на площині, що полегшує аналіз даних високої розмірності.

З огляду на актуальність візуалізації даних метою висвітленого у статті дослідження є подолання цього недоліку. Завданнями дослідження є розроблення методу візуалізації кластерної структури даних модельованих ДІСК і проведення аналізу адекватності цього методу для даних високої розмірності.

Зв’язок висвітленої проблеми із важливими науковими та практичними завданнями. Аналіз останніх публікацій і досліджень

Візуалізація кластерної структури даних високої розмірності є актуальним науковим і практичним завданням. Про це свідчать численні сучасні наукові публікації, в яких висвітлено не лише методи

візуалізації, але й їхнє застосування до розв'язання прикладних задач [19–24]. Зокрема, у праці [19] досліджено застосування керованих користувачами методів кластеризації з метою виявлення просторових залежностей у структурі генів. Авторами виділено візуалізацію як ключову техніку виявлення таких залежностей і прийняття рішення на основі візуальних спостережень за структурою кластерів. У праці [20] досліджено використання технології GHSOM (ієрархічної самоорганізовної карти з можливостями динамічного розростання) для аналізу трафіку даних з метою виявлення кримінальних дій (наприклад, здійснення інформаційних атак). Зауважено, що такий аналіз вимагає опрацювання величезних об'ємів даних і врахування варіацій самих атак. Розроблений метод візуалізації, що використовує GHSOM, забезпечує можливість спостереження ієрархічних залежностей даних. Це, на переконання авторів, забезпечує суттєво краще розуміння трафіку даних і спрощує пошук доказів про атаки.

У праці [21] наведено результати візуалізації класів даних, в основу якої покладено проектування результатів кластеризації у двовимірний простір. Автори розробили метод візуалізації на основі імовірнісного моделювання, яке відіграє регуляторну роль для випадку даних високої розмірності. Автори праці [23] розробили фреймворк аналізу даних за назвою Auto-HDS. Головною метою цього методу є автоматичний вибір кластерів даних на основі густини даних і відображання їхньої ієрархічної структури. Розроблений фрейворк забезпечує можливість візуалізації ієрархічної структури кластерів у двовимірному просторі. Автори наголошують на важливості візуалізації з метою поглибленого аналізу даних. У праці [24] проаналізовано 72 дільниці міста Мехіко на предмет виявлення закономірностей у даних про пересування транспорту, соціального забезпечення, нерухомості, якості повітря тощо. Сомоорганізовні карти Кохонена було використано з метою візуалізації структури даних, що уможливило виявлення нових закономірностей. Ці нові закономірності виявилися критично важливими для планування розвитку міста. Зокрема було отримано засіб аналізу географічного розподілу населення за демографічними та економічними показниками.

Аналіз отриманих наукових результатів

Моделі елементів карт Кохонена і ДІСК

Структурні особливості карти Кохонена і ДІСК є їхньою головною відмінністю з погляду візуалізації даних. Для кращого сприйняття розробленого методу візуалізації даних модельованих ДІСК у цьому підрозділі описано математичні моделі обох карт.

Елементи карти Кохонена

Головною функцією карти Кохонена є конвертування нелінійної статистичної залежності в даних великої розмірності у прості геометричні відношення елементів двовимірної ґратки. Самоорганізовною картою Кохонена називають нелінійне, впорядковане і гладке відображення $\Phi : X \to M$, де X – неперервний лінійний метричний простір великої розмірності; M – дискретний метричний двовимірний простір із топологією, яка визначається розташуванням його елементів у вузлах ґратки. Схематично відображення Φ, яке формується у результаті навчання карт Кохонена, зображено на рис. 1.

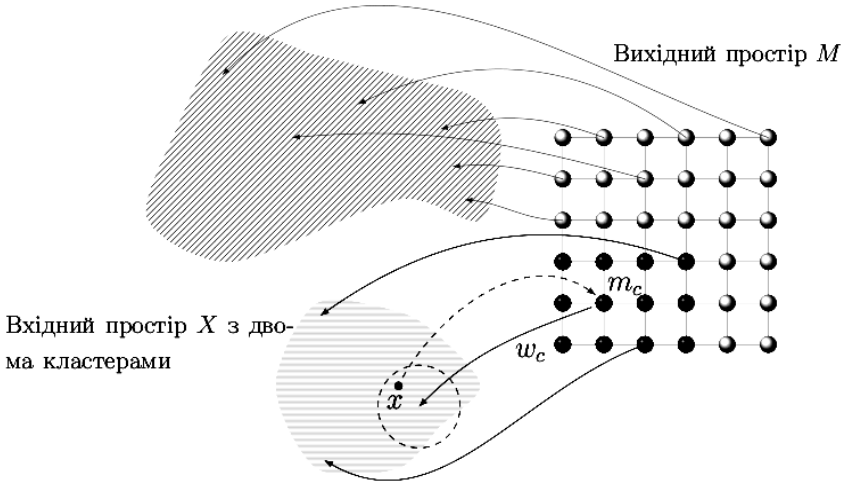


Рис. 1. Схематичне зображення апроксимації вхідного простору відображенням Ф

Елементи вихідного простору позначатимемо $m_k \in M$, $k = \overline{1,|M|}$, де $|M|$ – кількість елементів карти. Із кожним елементом m_k асоційований вектор $w_k \in W \subset X$ ($|M| = |W|$), який називають ваговим. Конфігурація ґратки визначається взаємним розміщенням елементів $m_k \in M$, які мають цілочислові координати і утворюють евклідів простір. Найпоширеніші схеми формування ґратки елементів – стільникова і прямокутна. Значення координат не є важливим – головне, щоб вони відповідали регулярній структурі ґратки. Кожен k-й елемент ґратки ототожнено з двома величинами – ваговим вектором w_k та цілочисловими координатами (i, j), які відповідають структурі та позиції елемента у ґратці. Позначимо цей факт як $m_k = \langle w_k, (i, j) \rangle$, $k = \overline{1,|M|}$.

Елементи ДІСК

Як і у випадку з картою Кохонена, множину елементів ДІСК позначимо M, а її елементи – $m_k \in M$, які називатимемо метанейронами. Кожен елемент $m_k \in M$ характеризуються трійкою величин $\langle W_k, S_k, w_k \rangle$. Тут $w_k = \left(w_k^1, \cdots, w_k^m\right)$ точковий ваговий вектор елемента m_k, де $m = |M|$. Величина $W_k = \left(W_k^1, \cdots, W_k^m\right) = \left([\overline{w_k^1}, \underline{w_k^1}], \cdots, [\overline{w_k^m}, \underline{w_k^m}] \right)$ називається інтервальним ваговим вектором, що визначає гіперкубічну область вхідного простору, за відображення даних з якої відповідає елемент m_k. Одним із можливих трактувань числового інтервалу є інтервал як множина чисел між його початком і кінцем [25]. На підставі такого трактування, $S_k = \left(S_k^1, \cdots, S_k^m\right)$ – вектор, компонентами якого є множини, відповідні інтервальним компонентам вектора W_k. З огляду на це інтервальний ваговий вектор можна записати як $W_k = \left([\min(S_k^1), \max(S_k^1)], \cdots, [\min(S_k^m), \max(S_k^m)] \right)$. Компоненти точкового вектора обчислюють як середні значення елементів множини вектора S_k, тобто $w_k = \left(\overline{S_k^1}, \cdots, \overline{S_k^m}\right)$.

Уведення інтервального вагового вектора забезпечує чітке визначення меж за кожним виміром вхідного простору, в яких елемент уважатиметься ефективним. Інтервальний ваговий вектор W_k не замінює точкового вагового вектора w_k, а відіграє винятково роль обмежувача. Спосіб побудови векторів W_k є основою навчального методу карти ДІСК [6, 16–18]. Геометричну інтерпретацію елемента ДІСК для випадку двовимірного дійснозначного вхідного простору подано на рис. 2.

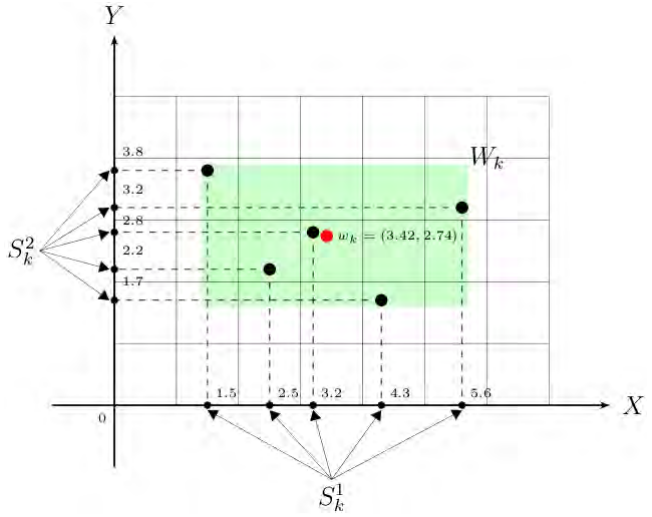


Рис. 2. Геометрична інтерпретація елементів ДІСК

Чорними кружальцями зображено вхідні дані $X = \{(1.5, 3.8), (2.5, 2.2), (3.2, 2.8), (4.3, 1.7), (5.6, 3.2)\}$, які моделює елемент m_k. Світло-сірим прямокутником зображено гіперкуб, який визна-

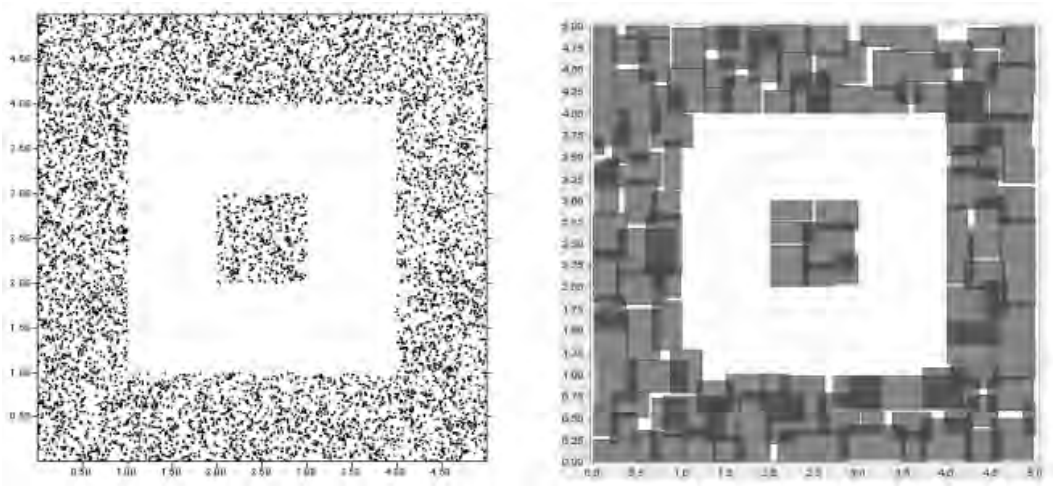
чається ваговим інтервальним вектором $W_k = ([1.5,5.6],[1.7,3.8])$. Сенсове навантаження компонент вектора $S_k = (\{1.5, 2.5, 3.2, 4.3, 5.6\},\{3.8, 2.2, 2.8, 1.7, 3.2\})$ полягає у відображенні проекцій векторів із вхідного простору на осі координат, за розпізнавання яких відповідає елемент m_k. Додатково на компоненти вектора S_k накладемо умову, що вони можуть містити повторювання значень. Це необхідно для коректного опрацювання випадків, коли проекції різних вхідних векторів на одну з осей збігаються. Сірим кружальцем посередині зображено точковий ваговий вектор w_k, координати якого обчислено як середні значення відповідних елементів вектора S_k.

Метод візуалізації даних модельованих ДІСК

Проектування елементів ДІСК у двовимірний простір

Відповідно до своїх структурних особливостей карта ДІСК не надає змоги візуалізувати кластерну структуру модельованих даних високої розмірності. Очевидною можливістю є лише випадок двовимірного вхідного простору даних.

Приклад такої візуалізації, отриманої для множин даних із двома лінійнонероздільними підмножинами, подано на рис. 3. Множина даних X, яку зображено на рис. 3(a), складається з двох вкладених підмножин, кожна з яких містить рівномірно розподілені елементи. Потужність множини $|X| = 6813$ точок. Як можна переконатися з рис. 3(b), навчена ДІСК розпізнала наявність обох підмножини і відобразила вкладену підмножину елементами, покриття яких не містить спільних точок зі зовнішньою підмножиною. Навчена на множині X ДІСК складалася зі 125 елементів.



(a) Вхідні дані (b) Модель даних ДІСК

Рис. 3. Вхідні дані (a), інтервальні вагові вектори елементів навченої ДІСК

Прямокутні області на рис. 3(b) зображають двовимірні інтервальні вектори ДІСК. Для даних вищої розмірності, ніж 2, гіперкубічні області, що відповідають інтервальним ваговим векторам елементів ДІСК, не є придатними для відображення на площині. І отже, виникає потреба проектування вагових векторів у двовимірний простір. З огляду на те, що карти Кохонена володіють двовимірною ґраткою, а елементи ДІСК – точковими ваговими векторами, за основу методу проектування елементів ДІСК було взято метод навчання карт Кохонена. Для цього усі точкові вагові вектори ДІСК трактують як множину вхідних даних X.

Карти Кохонена володіють трьома важливими властивостями, які є фундаментальними для якомога точнішого відображення *m*-вимірних вагових векторів елементів ДІСК у двовимірний простір.

1. *Апроксимація вхідного простору*. Відображення Φ, яке сформоване з множини вагових векторів W карти Кохонена у вихідному просторі M, забезпечує добру апроксимацію (у сенсі середньоквадратичного відхилення) вхідних даних X.

2. *Топологічне впорядкування.* Відображення Φ є топологічно впорядкованим у сенсі просторового розташування елементів карти, яке відповідає областям, що утворюють вектори вхідного простору. Тобто близьким векторам вхідного простору X (тут точкові вагові вектори елементів ДІСК) відповідають близькі елементи вихідного простору M (тут елементи двовимірної ґратки карти Кохонена). Близькість вхідних векторів визначається відповідно до метрики вхідного простору, а близькість елементів карти – у сенсі близькості їхніх координат (i, j) відповідно до метрики L_2.
3. *Відображення густини.* Області вхідного простору X, з яких вектори вибираються з більшою ймовірністю, відображаються більшою кількістю елементів вихідного простору M. Отже, області, вектори яких вибираються з більшою ймовірністю, відображені картою Кохонена детальніше, ніж області, вектори з яких вибираються з меншою ймовірністю.

Ці самі властивості має карта ДІСК [6]. Саме тому, трактуючи множину точкових вагових векторів елементів ДІСК як вхідну множину даних X для навчання карти Кохонена, отримуємо спосіб проектування елементів ДІСК у двовимірний простір зі збереженням їхнього топологічного впорядкування, яке відповідає топологічному впорядкуванню модельованого набору оригінальних даних.

Водночас цього недостатньо для візуалізації кластерної структури модельованих даних. Карта Кохонена, утворена в результаті навчання на елементах ДІСК, вимагає застосування до неї спеціалізованих алгоритмів візуалізації.

Візуалізація кластерної структури даних

Попри те, що завдяки своїм властивостям карта Кохонена містить усю необхідну інформацію про кластерну структуру модельованих даних, вона не володіє механізмом відтворення цієї інформації як зрозумілої людині, яка аналізує дані. Для цього використовують методи, які опрацьовують навчені карти Кохонена з метою виявлення структурної інформації у даних. Найпоширенішим методом для виявлення кластерної структури даних на основі карт Кохонена є U-Matrix [3] та низка його модифікацій [4, 5].

Суть методу U-Martix полягає у використанні двовимірної ґратки карти для побудови двовимірного растрового зображення структури модельованих даних, обчислюючи відстань між ваговими векторами сусідніх елементів карти. Тобто для кожного елемента карти $m \in M$ обчислюють значення, яке називають U-висотою, як відстань між ваговим вектором елемента w_k та ваговими векторами елементів його безпосереднього сусідства *M (m)*. А саме, $U_h(m) = \sum_{m' \in M(m)} d(w_m, w_{m'})$, де $d(w_m, w_{m'})$ – відстань між елементами ґратки $m, m' \in M$ у сенсі метрики вхідного простору даних.

Більше значення U_h вказує на більшу відстань між сусідніми елементами. Переважно елементу карти з найменшим значенням U_h відповідає піксел зображення білого кольору, а з найбільшим – чорного. Часто для точнішого визначення кластерних областей встановлюють деякий поріг для значень U_h. Усім елементам карти, які мають значення U_h, що не перевищує встановленого порогу, має відповідати піксел білого кольору. Для решти елементів кольори пікселів визначають як відтінок сірого кольору відповідно до величини значення U_h. Потім будують растрове зображення розміром, який відповідає розміру ґратки карти Кохонена, де пікселі мають координати і кольори елементів карти. Утворене зображення називають картою висот. Приклад карти висот подано на рис. 4.

Рис. 4. Приклад візуалізації кластерної структури даних за допомогою U-Matrix

На цьому рисунку зображено білі області, розділені границями, які зображено різними рівнями насиченості сірого кольору. Кожна біла область зображає окремий кластер даних. Інтенсивність забарвлення границь, які розділяють області кластерів, вказує на близькість цих кластерів між собою в сенсі метри вхідного простору даних: що інтенсивніше забарвлення, то більша відстань між двома кластерами.

З метою візуалізації кластерної структури даних саме із використанням моделі даних ДІСК необхідно застосовувати алгоритм U-Matrix не до вагових векторів навченої на елементах ДІСК карти Кохонена, а до точкових вагових векторів карти ДІСК. Водночас для побудови зображення пікселі визначені алгоритмом для елементів карти ДІСК, розташовуватимемо на двовимірній площині. Виникає запитання: як визначити, де саме повинен розташовуватися кожен піксел? Для цього скористаємося відображенням Φ^{-1}, оберненим до побудованого у процесі навчання карти Кохонена відображення Φ.

Відповідно до відношення Φ кожному вхідному вектору даних (у нашому випадку, точковому вектору елементів карти ДІСК) $x \in X$ відповідає елемент карти Кохонена, що виражається в обчисленні значення формули (1):

$$m(x) = \arg\min_{k=\overline{1,|W|}} d(x, w_k), \tag{1}$$

де $x \in X$ – вектор вхідного простору, для якого шукають елемент якнайкращого наближення, d – метрика вхідного простору, W – множина вагових векторів, $w_k \in W$ – ваговий вектор k-го елемента карти Кохонена, $m(x) \in M$ – шуканий елемент якнайкращого наближення, який належить множині M елементів карти Кохонена.

Для побудови оберненого відображення Φ^{-1} формулу (1) потрібно перетворити так, аби результатом її застосування був не елемент карти Кохонена, а елемент ДІСК. Формула (2) реалізує таке перетворення:

$$m'(w_k) = \arg\min_{j=\overline{1,|W'|}} d(w_k, w'_j). \tag{2}$$

Тут $w_k \in W$ – ваговий вектор k-го елемента карти Кохонена, для якого шукається елемент якнайкращого наближення карти ДІСК; W – множина вагових векторів карти Кохонена; d – метрика вхідного простору; W' – множина точкових вагових векторів елементів ДІСК; $w'_k \in W'$ – точковий ваговий вектор k-го елемента карти ДІСК; $m'(w_k) \in M'$ – шуканий елемент якнайкращого наближення, який належить множині M' елементів карти ДІСК.

Отже, кожному елементу карти ДІСК відображення Φ^{-1} ставить у відповідність множину елементів карти Кохонена. Для побудови карти висот, кожний елемент карти Кохонена на двовимірній ґратці зображається пікселем, колір якого обчислено для відповідного елемента карти ДІСК за алгоритмом U-Matrix.

Аналіз результатів візуалізації кластерної структури даних модельованих ДІСК

Для перевірки адекватності розробленого методу візуалізації ДІСК було проведено низку експериментів на еталонних наборах даних, використання яких є загальноприйнятою практикою [26]. З метою подальшого обговорення обрано набір даних Zoo, який містить інформацію про 101 живу істоту, кожну з яких описано вектором 18-ти ознак: назва істоти, номер кластера (ознака прийняття рішення), 15-ть двійкових ознак і одна цілочислова. Фрагмент цих даних подано у табл. 1, де заголовки стовпців відповідають назвам ознак. Сформована для побудова моделі за допомогою карти ДІСК множина даних X містила вектори розмірності 16 – усі ознаки, окрім назви тварин (ознака Name) і номера кластера (ознака Type).

Таблиця 1

Фрагмент вхідних даних Zoo

Name	Hair	Feathers	Eggs	Milk	Airborne	Aquatic	Predator	Toothed	Backbone	Breathes	Venomous	Fins	Legs	Tail	Domestic	Catsize	Type
aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	1
antelope	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	1
bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
bear	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	1
boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	1
buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	1
calf	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	1

У результаті процесу навчання карти ДІСК, який було реалізовано відповідно до алгоритму навчання, описаному в праці [6], отримано 11 елементів карти. Результат розпізнавання кожного з вхідних векторів подано у табл. 2. Перший стовпець у цій таблиці містить значення ознаки прийняття рішення, тобто наперед відоме значення кластера, до якого належать перелічені тварини. Другий стовпець містить номер елемента навченої ДІСК, який відповідає за розпізнавання підмножини тварин. Третій стовпець містить підмножини вхідної множини даних, які моделюються відповідним елементом карти ДІСК. Для зручності вхідні дані подано відповідними назвами тварин замість векторного подання їхніх ознак, що використовувались для навчання.

Таблиця 2

Належність даних до кластерів й елементів ДІСК

Номер кластера	Елементи ДІСК	Дані
1	6	antelope, buffalo, deer, elephant, giraffe, oryx, calf, goat, pony, reindeer, cavy, pussycat, mole, opossum, hamster, hare, vole, dolphin, porpoise, seal, sealion, fruitbat, vampire, squirrel, wallaby
	7	aardvark, bear, boar, cheetah, leopard, lion, lynx, mongoose, polecat, puma, raccoon, wolf
	8	girl, gorilla
2	4	chicken, dove, duck, flamingo, gull, lark, ostrich, parakeet, penguin, pheasant, rhea, skimmer, skua, sparrow, swan, vulture, wren
	5	crow, hawk, kiwi
3	2	pitviper, seasnake, slowworm, tortoise, tuatara
4	3	bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, sole, stingray, tuna
5	9	frog, frog, newt, toad
6	10	honeybee, housefly, moth, wasp
	11	flea, gnat, ladybird, termite
7	1	clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm

Із табл. 2 видно, що тварини з першого кластера моделюються трьома елементами 6, 7 і 8; тварини з другого кластера – елементами 4 і 5; третього – елементом 2 тощо. Тобто за розпізнавання даних із деяких кластерів відповідають декілька елементів карти ДІСК, а із деяких – по одному елементу. Проаналізувавши для прикладу перший кластер, легко переконатися, що його три підмножини, елементи яких розпізнаються трьома різними елементами карти ДІСК, справді утворюють окремі підкластери. Елемент 6 моделює нехижих (переважно травоїдних) тварин, елемент 7 – хижих (окрім aardvark (мурахоїд) і boar (вепр), які за більшістю перелічених ознак потрапляють до

хижих), а елемент 8 – людиноподібних. Схожого висновку про існування підкластерів можна дійти, аналізуючи дані, які розпізнаються елементи 4 і 5 (кластер 2), 10 і 11 (кластер 6).

Застосовуючи розроблений метод візуалізації до навченої карти ДІСК, було отримано карту висот, яка зображена на рис. 5. Білі області зображають окремі групи близьких даних і є промарковані номером елемента карти ДІСК, який відповідає за розпізнавання даних із відповідного кластера. Межі між кластерами зображено лініями сірого кольору, насиченість яких вказує на більшу (темніший відтінок), або меншу (світліший відтінок) віддаленість суміжних кластерів.

Аналізуючи зображення на рис. 5, легко переконатися, що вищеподані міркування про існування підкластерів у кластерах, які визначені на основі експертної оцінки, є також застосовними й до карти висот. Як і перед тим, звернемо увагу на інтенсивність кольору границь, що відмежовують області, промарковані номерами елементів 7 і 8 від сусідньої області з номером 6. Це є світло-сірі границі, які вказують на те, що модельовані відповідними елементами ДІСК дані є ближчими до даних, модельованих елементом 6, ніж, скажімо, елементом 9 чи 4, і належать до одного кластера. Схожу ситуацію можна спостерігати із областями, які відповідають елементу 5 – підкластер для елемента 4, і 11 – підкластер для елемента 10. Об’єднання цих областей і перенумерування решти присвоєнням відповідних номерів кластерів дає можливість побудови карти висот, яку зображено на рис. 6.

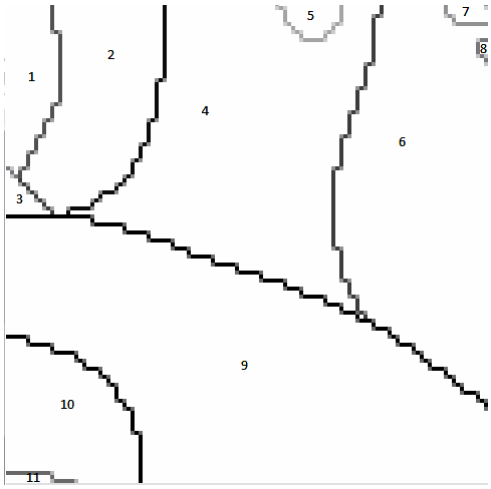


Рис. 5. Карта висот для ДІСК, навченої на наборі даних Zoo. Номери вказують на елементи ДІСК

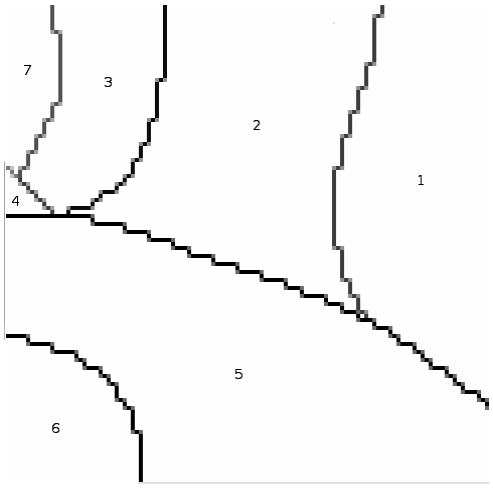


Рис. 6. Модифікована карта висот для ДІСК, навченої на наборі даних Zoo. Номери вказують на кластери, встановлені на основі експертної оцінки

Окремо слід звернути увагу на той факт, що границі між суміжними кластерами 3, 4 і 7 (рис. 6) забарвлені у дещо світліший колір порівняно з границями із кластерами 2 та 5. Аналізуючи дані у табл. 2, можна переконатися, що тварини з кластерів 3, 4 і 7 й справді близькі за своєю природою – інформація, яка не є відразу очевидною. Саме завдяки природному (тобто використовуючи дані без жодного попереднього експертного оцінювання) групуванню даних карти Кохонена і ДІСК дають змогу відслідковувати і візуалізовувати такі приховані залежності.

Висновки і перспективи подальших наукових досліджень

Розроблений метод дав змогу зробити якісний крок у напрямку вирішення питання візуалізації даних модельованих ДІСК. Результати експериментів вказують на адекватну візуалізацію кластерної структури даних високої розмірності. Водночас, схоже як і для карт Кохонена [27], розроблений метод не володіє здатністю автоматичного об’єднання дрібних кластерів в укрупнені кластери, що ускладнює аналіз даних. Для демонстрації цього недоліку скористаємося набором даних Iris, для якого побудовано карту висот ДІСК. На рис. 7 зображено отриману карту.

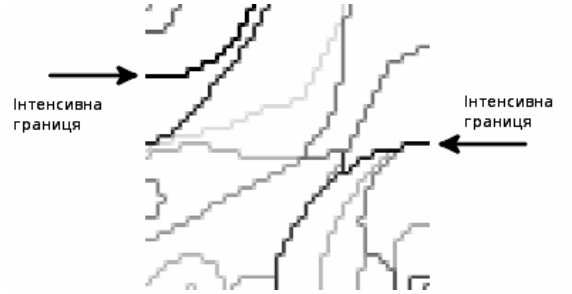


Рис. 7. Карта висот ДІСК для набору даних Iris

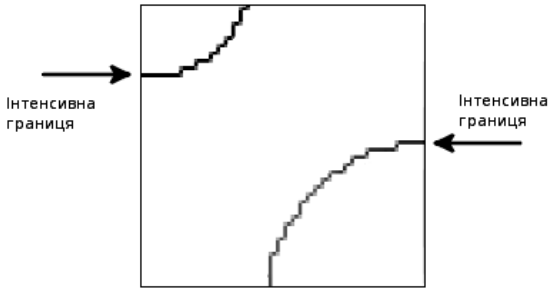


Рис. 8. Модифікована карта висот ДІСК для набору даних Iris

Легко переконатися, що використовуючи лише це зображення, доволі складно визначити, скільки кластерів містять дані. Знаючи, що набір даних Iris містить три кластери, на карті висот можна виділити дві границі найінтенсивнішого кольору. Справді, ці границі розмежовують три кластери даних. Бажаний результат візуалізації подано на рис. 8, де залишено лише ці дві границі. У наших попередніх дослідженнях було розроблено напівавтоматичний метод укрупнення кластерів даних для карти Кохонена [27]. У подальших дослідженнях планується апробація цього методу до карт ДІСК.

Додатково, слід звернути увагу на час виконання розробленого методу візуалізації ДІСК порівняно із технологією карт Кохонена. На рис. 9 подано час навчання карт Кохонена на чотирьох еталонних наборах даних (заштрихована гістограма) і сумарний час, який було затрачено для навчання ДІСК на тих самих наборах даних, плюс час для навчання карт Кохонена на множинах точкових вагових векторів ДІСК. З огляду на особливості навчального алгоритму карти ДІСК час отримання карт Кохонена для проектування елементів ДІСК у двовимірний простір для усіх використаних наборів даних був утричі швидший, ніж побудова карти Кохонена на оригінальних наборах даних.

Рис. 9. Час навчання карти Кохонена на оригінальних даних порівняно із сумарним часом навчання ДІСК на цих самих даних і карти Кохонена на вагових векторах елементів ДІСК

Зауважимо, що нами не проводилося асимптотичного оцінювання алгоритмічної складності методів навчання ДІСК. Це є одним із важливих завдань подальших досліджень і розвитку технології аналізу даних на основі ДІСК.

1. Dunham M. H. Data Mining: Introductory and Advanced Topics / M. H. Dunham. – Prentice Hall, 2003. – 315 p. 2. Pyle D. Data Preparation for Data Mining / D. Pyle. – Academic Press, 1999. – 540 p. 3. Ultsch A. Self-organizing neural networks for knowledge acquisition / A. Ultsch // Proceedings of

the 10th ECAI. – Vienna, Austria: 1992. – P. 208–210. 4. Ultsch A. Maps for the Visualization of high-dimensional Data Spaces / A. Ultsch // Proceedings of Workshop on Self Organizing Maps (WSOM03). – Kyushu, Japan: 2003. – P. 225–230. 5. Ultsch A. U-Matrix: a Tool to visualize Clusters in high dimensional Data / A. Ultsch // Technical Report. – University of Marburg, Department of Computer Science, 2003. – Vol. 36. – P. 1–12. 6. Годич О. В. Індуктивні методи та алгоритми самоорганізації моделей даних на основі карт Кохонена: дис. … канд. техн. наук : 01.05.03 / О.В. Годич. – Л., 2010. – 171 с. 7. Bauer H. Growing a hypercubical output space in a self-organizing feature map / H. Bauer, T. Villman // IEEE Transactions on Neural Networks. – 1997. – Vol. 8. – P. 218–226. 8. Blackmore J. Visualizing high-dimensional structure with the incremental grid growing neural network: Ph.D. thesis [Електронний ресурс] / The University of Texas at Austin. – 1995, Режим доступу: www.cs.utexas.edu/users/nn/downloads/ papers/blackmore.thesis.pdf. 9. Fritzke B. Growing cell structure – a self-organising network for unsupervised and supervised learning / B. Fritzke // Neural Networks. – 1995. – Vol. 8, no. 9. – P. 1441–1460. 10. Harp S.A. Genetic optimization of self-organizing feature maps / S.A. Harp, T. Samad // Proceedings of IEEE International Joint Confecence on Neural Networks. – Vol. 1. – Seattle, WA, USA: 1991. – P. 341–346. 11. Huang S. J. Genetic algorithms enhanced kohonen's neural networks / S. J. Huang, C. C. Hung // Proceedings of IEEE International Joint Conference on Neural Networks. – Vol. 2. – USA: 1995. – P. 708–712. 12. Koh J. A multilayer self-organizing feature map for range image segmentation / J. Koh, M. Suk, S. M. Bhandarkar // Neural Networks. – 1995. – Vol. 8, No. 1. – P. 67–86. 13. Si J. Dynamic topology representing networks / J. Si, S. Lin, M. Vuong // Neural Networks. – 2000. – Vol. 13, No. 6. – P. 617–627. 14. Su M. C. Genetic-algorithm-based approach to self-organizing feature map and its application in cluster analysis / M. C. Su, H. T. Chang // Proceedings of IEEE International Joint Conference on Neural Networks. – Alaska, USA: 1998. – P. 2116–2121. 15. Загоруйко Н. Прикладные методы анализа данных и знаний / Н. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. – 270 с. 16. Годич О. Кластеризація даних нейромережею ADD / О. Годич // Вісник НУЛП, Інформаційні системи та мережі. – 2005. – No 549. – С. 54–68. 17. Годич О. Динамічна нейромережа ADD / О. Годич, Ю. Щербина // Вісник Львівського нац. ун-ту ім. І.Франка, серія "Прикладна мат. та інформ.". – 2005. – No 10. – С. 161–183. 18. 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. 19. Rubel O. Integrating Data Clustering and Visualization for the Analysis of 3D Gene Expression Data , O. Rubel et al. // IEEE/ACM Trans. Comput. Biol. Bioinformatics. – 2010. – Vol. 7, No. 1. – P. 64–79. 20. Palomo E. J. 2012 Special Issue: Application of growing hierarchical SOM for visualisation of network forensics traffic data. / E. J. Palomo et al. // Neural Netw. – 2012. – Vol. 12. – P. 275–284. 21. Kaban A. On class visualisation for high dimensional data: exploring scientific data sets. / A. Kaban et al.// Proceedings of the 9th international conference on Discovery Science (DS'06). – 2006. – P. 125–136. 22. Lisboa P. J. G. Cluster-based visualisation with scatter matrices. / Lisboa, P. J. G. et al. // Pattern Recogn. Lett. – 2008. – Vol. 29, No. 13. – P. 1814–1823. 23. Gupta G. Automated Hierarchical Density Shaving: A Robust Automated Clustering and Visualization Framework for Large Biological Data Sets. / G. Gupta, A. Liu, J Ghosh // IEEE/ACM Trans. Comput. Biol. Bioinformatics. – 2010. – Vol. 7, No. 2. – P. 223–237. 24. Neme O. Mining the city data: making sense of cities with self-organizing maps / Neme O, Pulido J., Neme A. // Proceedings of the 8th international conference on Advances in self-organizing maps (WSOM'11). – 2010. – P. 168–177. 25. Hansen E. Global Optimization Using Interval Analysis: Revised And Expanded / E. Hansen, G. W. Walster. – 2nd edition. – CRC Press, 2003. – 728 p. 26. Asuncion A. UCI machine learning repository.– 2007, Режим доступу: http://archive.ics.uci.edu/ml. 27. Hodych O. Determining cluster boundaries within Self-Organizing Maps / O. Hodych, I. Nikolski, V. Pasichnyk, Y. Shcherbyna // Вісник ХПІ. – 2007. – No 5. – С. 97–109.*