Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Метод кластеризації даних з використанням машинного навчання

(тема)

Виконав:

студент II курсу, групи СПзм-20-1

Хануков П.Д.
(прізвище, ініціали)

Спеціальність

123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма

Системне програмування
(повна назва освітньої програми)

Керівник: доц. Ткачов В.М.
(посада, прізвище, ініціали)

Допускається до захисту

В.о. зав. кафедри ЕОМ _____ Волк М.О.
(підпис) (прізвище, ініціали)

2022 р.

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Ханукову Павлу Дмитровичу
(прізвище, ім'я, по батькові)

1. Тема роботи Метод кластаризації даних з використанням методів машинного навчання

затверджена наказом по університету від "08" квітня 2022 р. № 52 Стз

2. Термін подання студентом роботи до екзаменаційної комісії 18 травня 2022 р.

3. Вхідні дані до роботи
машинне навчання
штучні нейронні мережі
карти Кохонена

4. Перелік питань, що потрібно опрацювати у роботі
Карти Кохонена як метод машинного навчання
Моделі модифікованих карт Кохонена з замкнутою решіткою
Моделювання модифікованої мережі Кохонена

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 14 слайдів

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналіз предметної області	11.04.2022–20.04.2022	
2	Дослідження апарату карт Кохонена	21.04.2022–26.04.2022	
3	Розробка методу кластеризації	27.04.2022–29.04.2022	
4	Моделювання та перевірка роботи моделі	30.04.2022–02.05.2022	
5	Отримання та аналіз результатів	03.05.2022–06.05.2022	
6	Оформлення пояснювальної записки	07.05.2022–13.05.2022	

Дата видачі завдання 11 квітня 2022 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Ткачов В.М.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 81 с., 27 рис., 4 табл., 1 дод., 6 джерел.

КАРТА КОХОНЕНА, НЕЙРОННА МЕРЕЖА, МАШИННЕ НАВЧАННЯ, КЛАСТЕРИЗАЦІЯ.

Метою кваліфікаційної роботи є дослідження методів кластеризації даних за допомогою методів машинного навчання, зокрема карт Кохонена.

У ході виконання кваліфікаційної роботи були вивчені карти Кохонена, що самоорганізуються. Розглянуті як класичні, так і модифіковані варіанти навчання мережі. Розроблене програмне забезпечення, що реалізує дані алгоритми. Проведені дослідження ефективності навчання модифікованих мереж Кохонена. Найкращі результати показав метод навчання мережі Кохонена з замкнутою решіткою.

ABSTRACT

Master’s thesis: 81 pages, 27 figures, 4 tables, 1 appendices, 6 sources.

KOHONEN MAP, NEURAL NETWORK, MACHINE LEARNING, CLUSTERING.

The major goal of this thesis is to study the methods of data clustering using machine learning methods, in particular Kohonen maps. Self-organizing Kohonen maps were studied. Both classical and modified variants of network training are considered. Developed software that implements these algorithms. Studies of the effectiveness of learning Kohonen's modified networks have been conducted. The method of training the Kohonen network with a closed lattice showed the best results.

In order to Self-organizing Kohonen maps were studied. Both classical and modified variants of network training are considered. Developed software that implements these algorithms. Studies of the effectiveness of learning Kohonen's modified networks have been conducted. The method of training the Kohonen network with a closed lattice showed the best results.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ 7
ВСТУП 8
1 КАРТИ КОХОНЕНА ЯК МЕТОД МАШИННОГО НАВЧАННЯ 11
1.1 Аналіз сучасних методів машинного навчання для обробки даних 11
1.2 Класична модель СОКК і алгоритм її навчання 21
1.3 Підготовка і попередня обробка даних для нейронної мережі 33
1.4 Побудова візуальних топографічних карт для СОКК 36
2 МЕТОД ВДОСКОНАЛЕННЯ АПАРАТУ КАРТ КОХОНЕНА 43
2.1 Проблеми застосування класичної моделі СОКК 43
2.2 Способи усунення граничного ефекту 44
2.3 Новий метод зв'язку сусідніх нейронів мережі 48
2.4 Нові моделі СОКК із замкнутими гратами для усунення граничного ефекту 52
3 МОДЕЛЮВАННЯ МОДИФІКОВАНОЇ МЕРЕЖІ КОХОНЕНА 58
3.1 Оцінка точності і якості навчання мережі із замкнутими решітками 58
3.2 Ентропія мережі як оцінка якості навчання 61
3.3 Вибір інструментальних засобів 67
ВИСНОВКИ 72
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ 73
ДОДАТОК А Графічний матеріал кваліфікаційної роботи 74

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

АІС – автоматизована інформаційна система

ПК – програмний комплекс

СМК – система менеджменту якості

СОД – система обробки даних

СОКК – самоорганізована карта Кохонена

СППР – система підтримки прийняття рішень

СКБД – система керування базами даних

ВСТУП

Сучасний рівень розвитку апаратних та програмних засобів з деяких пір уможливив повсюдне ведення баз даних оперативної інформації на всіх рівнях управління. В процесі своєї діяльності промислові підприємства, корпорації, органи державної влади і управління накопичили великі обсяги різнорідних даних. Ці дані зберігають у собі великі потенційні можливості по вилученню корисної аналітичної інформації, на основі якої можна виявляти приховані тенденції, будувати стратегію розвитку, знаходити нові рішення.

В останні роки використовуються наступні концепції накопичення і аналізу даних:

- сховища даних (англ.: Data Warehouse) [1, 2, 3];
- оперативна аналітична обробка даних (англ.: On-Line Analytical Processing, OLAP) [4];
- інтелектуальний аналіз даних (англ.: Data Mining) [5].

Метою побудови сховища даних є інтеграція, актуалізація і узгодження оперативних даних з різнорідних джерел для формування єдиного несуперечливого погляду на об'єкт управління в цілому.

Завданнями оперативної аналітичної обробки даних є узагальнення, агрегація, багатовимірний аналіз і гиперкубической уявлення інформації, зібраної в сховище даних.

Інтелектуальний аналіз даних – це процес підтримки прийняття рішень, заснований на пошуку в даних схованих закономірностей.

Системи обробки даних розділяються на два класи:

- системи, орієнтовані на трансакціонної обробку даних (СОД);
- системи підтримки прийняття рішень (СППР).

СОД призначені для розв'язання добре структурованих задач, за якими є вхідні дані, відомі алгоритми, що ведуть до вирішення завдань. Система працює з мінімальною участю людини. Основними функціями СОД є збір

даних і перенесення їх на машинні носії, передача в місця зберігання і обробки, зберігання, обробка інформації за стандартними алгоритмами, висновок та подання інформації користувачу у вигляді регламентних форм [6].

СППР використовуються для вирішення в режимі діалогу погано структурованих задач, для яких характерна неповнота вхідних даних, часткова ясність цілей і обмежень. Участь людини в роботі системи велике, він може втручатися в хід рішення, модифікувати вхідні дані, процедури обробки, цілі та обмеження завдання, вибирати стратегії оцінки варіантів рішень. СППР використовується на рівні стратегічного планування, оперативного і управлінського контролю [6].

Дана кваліфікаційна робота присвячена питанням підвищення ефективності прийняття рішень з використанням сучасних методів інтелектуального аналізу даних в СППР.

За допомогою сучасних методів інтелектуального аналізу даних можна вирішуються наступні категорії задач :

а) завдання класифікації. Використання методів дозволяє виявити ознаки, що характеризують однотипні групи об'єктів - класи, - для того щоб за відомими значеннями цих характеристик можна було віднести новий об'єкт до того чи іншого класу. Методами вирішення завдання класифікації є: метод к-середніх , Байесови мережі, дерева рішень, нейронні мережі;

б) завдання кластеризації. Логічно продовжують ідею класифікації на більш складний випадок, коли самі класи заздалегідь не визначені. Результатом використання методів, що виконують кластеризацию, являється визначення тих властивих досліджуваним даними ознак, які розбивають їх на групи. Такі групи даних називаються кластерами. Об'єкти даних усередині кластерів "схожі" один на одного і "відрізняються" від об'єктів даних інших кластерів. Методами вирішення завдання кластеризації є нейронні мережі, навчають без вчителя – само організуються карти Кохонена;

в) завдання асоціації. Методи виконує пошук асоціацій не на основі

значень властивостей одного об'єкта або події, а між двома або декількома одночасно наступаючими подіями. При цьому вироблені правила вказують на те, що при настанні однієї події з тим або іншим ступенем ймовірності настає інший;

г) завдання виявлення шаблонів послідовностей. Подібно асоціаціям, послідовності мають місце між подіями, але наступаючими неодночасно, а з деяким певним розривом у часу. Таким чином, асоціація є окремий випадок шаблону послідовності з нульовою тимчасовою затримкою;

д) завдання прогнозування. Методи виконують оцінку певних майбутніх численних показників системи на основі особливостей поведінки поточних і історичних даних. Методами вирішення завдання є методи отримання статистичних закономірностей, методи регресійного і кореляційного аналізу;

У кваліфікаційній роботі використовуються сучасні методи машинного навчання, які вирішують завдання кластеризації, а саме я нейронні мережі Кохонена.

Метою кваліфікаційної роботи є дослідження методів кластеризації даних за допомогою методів машинного навчання, зокрема карт Кохонена.

Для досягнення поставленої мети були вирішені наступні наукові і практичні завдання:

- дослідження методів кластеризації даних;
- дослідження методів машинного навчання;
- дослідження алгоритму роботи класичної моделі нейронної мережі Кохонена, його недоліків і існуючих способів їх усунення;
- розробка методу удосконалення апарату нейронної мережі Кохонена для усунення недоліків мереж цього типу: граничного ефекту і появи "мертвих" нейронів;
- проведення моделювання з використанням розробленого методу.

1 КАРТИ КОХОНЕНА ЯК МЕТОД МАШИННОГО НАВЧАННЯ

У цьому розділі наводиться аналіз сучасних методів машинного навчання, зокрема штучних нейронних мереж, описується класична модель СОКК і алгоритм її навчання. Розглядаються основні елементи моделі, такі як, мережа нейронів, навчальна множина, алгоритм навчання. Розглядаються методи попередньої підготовки даних для аналізу за допомогою СОКК, а також способи побудови візуальних топографічних карт мережі.

1.1 Аналіз сучасних методів машинного навчання для обробки даних

Інтелектуальні інформаційні системи, що використовують сучасні методи інтелектуального аналізу даних, проникають в усі сфери нашого життя, тому важко провести сувору класифікацію напрямків, за якими ведуться активні і численні дослідження в області штучного інтелекту.

Самоорганізовані інтелектуальні системи засновані на методах автоматичної класифікації ситуацій з реальної практики, або на методах навчання на прикладах. Приклади реальних ситуацій складають так звану навчальну вибірку, яка формується протягом певного історичного періоду. Елементи навчальної вибірки описуються безліччю класифікаційних ознак.

Стратегія "навчання з учителем" передбачає завдання фахівцем для кожного прикладу значень ознак, що показують його приналежність до певного класу ситуацій. При навчанні "без учителя» система повинна самостійно виділяти класи ситуацій за ступенем близькості значень класифікаційних ознак.

В процесі навчання проводиться автоматична побудова узагальнюючих правил або функцій, що описують приналежність ситуацій до класів, якими система згодом буде користуватися при інтерпретації незнайомих ситуацій. З узагальнюючих правил, в свою чергу, автоматично формується база знань,

яка періодично коригується в міру накопичення інформації про аналізовані ситуаціях.

У загальному випадку, основна ціль апарату штучних нейронних мереж ґрунтується, по-перше, на розпаралелюванні обробки інформації, по-друге, на здатності в отриманні обґрунтованого результату на підставі даних, які не зустрічалися в процесі навчання. Нейронні мережі працюють з неповними даними, тобто даними, в яких пропущені значення деяких класифікаційних ознак. Кількість обчислювальних операцій в них зростає лінійно зі збільшенням обсягу аналізується. Нейронні мережі не чутливі до розмірності аналізованих даних.

З огляду на класифікацію завдань, що вирішуються за допомогою сучасних методів інтелектуального аналізу даних, описані у введенні до даної атестаційної роботи, наведений аналіз цих методів.

Самоорганізовані карти Кохонена (мережі Кохонена, СОКК) – це штучні нейронні мережі, відмітною особливістю яких є автоматичний, некерований процес роботи з даними, що називається навчанням. При навчанні СОКК вчиться розуміти саму структуру даних, дає можливість користувачеві розпізнати кластери в даних. Якщо після навчання мережа зустрінеться зі спостереженнями, несхожими ні на один з відомих їй зразків, то вона не зможе класифікувати такий набір і тим самим виявить його новизну, тобто новий кластер даних.

Штучні нейронні мережі. Віднесення об'єктів, представлених векторами своїх ознак, до одного з заздалегідь відомих класів. Два етапи роботи алгоритму: на першому етапі відбувається навчання нейронної мережі шляхом пред'явлення їй всіх об'єктів і зазначенням для кожного приналежність до того чи іншого класу; на другому етапі мережу автоматично відносить пред'являється об'єкт до одного з класів, які вона була навчена розпізнавати на першому етапі.

Об'єкти вхідних даних є векторами, кожен з яких містить однакову кількість ознак. Нейронна мережу може складатися з одного нейрона -

персептрон, а може складатися з декількох нейронів. Кілька нейронів представляють собою шар нейронів. Нейронні мережі можуть складатися з декількох шарів, де виходи одного шару нейронів, є входами для іншого - наступного шару.

Переваги:

- висока точність;

- нейронна мережа нечутлива до пропусків і шумів (викидам) в значеннях і ознаках спостережень, присутніх як в навчальній вибірці, так і в довільному об'єкті.

Недоліки:

- навчання з учителем;

- необхідна велика навчальна вибірка спостережень за об'єктом, щоб натренувати мережу вьшолняют класифікацію. Чим більше спостережень, тим точніше класифікація мережею.

К-найближчих сусідів. Автоматична класифікація об'єктів. Кожен об'єкт буде віднесений до того класу, який переважає серед до найближчих об'єктів-сусідів поточного об'єкта.

Вибирається к. Далі для кожного вектора знаходяться до векторів, з мінімальним евклідової відстанню до нього. Після цього кожен вектор відноситься до того класу об'єктів, кількість яких переважає серед до знайдених векторів-сусідів.

Переваги:

- швидкість і простота реалізації.

Недоліки:

- проблема вибору к. Якщо до = 1 алгоритм дає помилкові класифікації. Якщо до дорівнює кількості об'єктів, то алгоритм надмірно стійкий і дає одну і ту ж классисифікацію.

Дерева рішень. Автоматичний аналіз даних, віднесення об'єктів до одного з заздалегідь відомих класів. Побудова деревовидної конструкції альтернатив типу "якщо ... то ...". Кожна альтернатива дає єдине правильне

рішення.

Вхідний набір даних складається з векторів, кожен з яких характеризується однаковим набором атрибутів. один ознака в кожному векторі показує, до якого класу вектор відноситься. Спочатку вхідний набір векторів містить вектори різних класів. Дерево розіб'є його на кілька підмножин. Для цього буде знайдений інший ознака (чи не класове), значення якого в різних векторах різний. Кількість записів, яку у цієї ознаки, на стільки підмножин буде розбитий вхідний набір даних. З отриманими підмножинами даних дерево надійде точно також: розіб'є їх на нові підмножини. такий рекурсивний процес буде тривати до тих пір, поки в одному підмножині не залишиться векторів, що належать одному класу.

Переваги:

- швидкий процес навчання;
- висока точність.

Недоліки:

- навчання з учителем: необхідність явного зазначення ознаки, по якому буде відбуватися класифікація об'єктів.

Векторне квантування. Поділ простору вхідних даних на класи, заміною цього простору кодують векторами.

Заздалегідь передбачається, що всі вектори вхідних даних взяті з кінцевого набору класів, які можуть перекриватися один з другом. Кожному з класів приписується деякий підмножина кодують векторів. Потім для кожного вектора вхідних даних відшукується свій кодує вектор, евклідова відстань до якого мінімально. Тоді вектор вхідних даних буде належати тому класу, якому належить цей кодує вектор.

Недоліки:

- складність визначення метрики, за допомогою якої оцінюється відстань між векторами даних;
- навчання з учителем.

Нечіткі дерева рішень. Автоматичний аналіз даних за допомогою

механізму дерев рішень, але приналежність об'єкта до якого-небудь Класу не повна, а має свою ступінь. За допомогою апарату нечіткої логіки визначається ступінь приналежності об'єктів до заданих класах. Рекурсивне розбиття вхідного простору даних на підмножини проводиться не за об'єктами, а за ступенем їх приналежності.

Переваги:

- висока точність;
- швидкий процес навчання;
- легко інтерпретується результат.

Недоліки:

- необхідний великий набір навчальних вхідних даних.

Байєсівський класифікатор. Визначення ймовірності приналежності об'єкта до одного з відомих класів.

Відомо, що об'єкт може знаходитися в одному з декількох станів (класів). На підставі безлічі спостережень (так званої навчальної виборки) за об'єктом обчислюються функції правдоподібності кожного класу для кожного спостереження об'єкта. Завдання методу: побудувати алгоритм, здатний класифікувати довільний об'єкт по одному з відомих класів.

Переваги:

- висока точність.

Недоліки:

- складність у відновленні щільності класів за навчальною вибірці для побудови алгоритму класифікації;
- необхідна навчальна вибірка великого розміру;
- апріорно приймається, що класи незалежні.

Нечіткі нейронні мережі. Використовується апарат нейронних мереж, але для розрахунку вихідного сигналу нейронів використовуються нечіткі правила, такі нейронні мережі складаються з декількох шарів. Перший шар мережі перетворює вхідні вектори в нечіткі множини (процедура фуззіфікації), що характеризуються кожне своєю функцією приналежності,

другий шар – агрегування значень активації умови, третій (лінійний) - агрегування заданого числа нечітких правил виводу (перший нейрон) і генерацію нормалізує сигналу (другий нейрон), четвертий шар – виконує нормалізацію, формуючи вихідний сигнал.

Переваги:

- висока точність класифікації;
- навчання, як з учителем, так і без вчителя.

Недоліки:

- зростає складність реалізації за рахунок застосування апараті нечіткої логіки.

Статистичні методи: логістична регресія. Передбачення ймовірності настання якої-небудь події (Подія відбудеться або подія не відбудеться) за значеннями безлічі ознак. При цьому на ймовірність настання події впливає все безліч ознак в тій або іншій мірі. Ступінь впливу кожної ознаки задається його коефіцієнтом. Для побудови регресійної моделі значення цих коефіцієнтів (коефіцієнтів регресії) визначаються за навчальною вибіркою. У навчальній вибірці для кожної відомої комбінації значень ознак відомий факт настання події (подія відбулася або не відбулося). Зв'язок між ймовірністю настання події і всіма впливають на нього ознаками підганяється під логістичну функцію.

Недоліки:

- обов'язкова наявність великої навчальної вибірки для підбору коефіцієнтів регресії.

Ієрархічна кластеризація: агломеративний алгоритм. Створюється ієрархія кластерів у вигляді дерева – дендограмм. Корінь дерева складається з одного кластера, що містить всі спостереження, а листя відповідають індивідуальним спостереженнями. Критерієм поділу або об'єднання спостережень між кластерами може бути, наприклад, функція попарних відстаней між спостереженнями.

Переваги:

- висока якість кластеризації;
- не вимагають апріорного вказівки кількості кластерів.

К-середніх. Автоматична кластеризация об'єктів на заздалегідь відоме кількість кластерів к. Випадково обрані до об'єктів визнаються центрами кластерів. Решта об'єктів приписуються до цих кластерів, якщо знаходяться до них ближче, ніж до інших. Серед об'єктів кожного кластера відшукується новий центр кластера і знову інші об'єкти розподіляються по кластерам вже нових центрів. Після того як центри кластерів перестають змінюватися, тобто закріплюються за одними і тими ж об'єктами, алгоритм зупиняється.

Переваги:

- швидкість і простота реалізації.

Недоліки:

- кількість кластерів об'єктів встановлюється заздалегідь як константа.

Само організована мережа Кохонена. Автоматична кластеризація і візуалізація багатовимірних даних. Мережа складається з нейронів. Нейрони розташовуються в вузлах двовимірної решітки. Кожен нейрон має свій ваговий вектор. Вектори з вхідного набору даних пред'являються мережі у випадковому порядку. При цьому для кожного з них відшукується ваговій вектор будь-якого нейрона, евклідова відстань до якого від вхідного вектора мінімально. Такий ваговій вектор і вагові вектори нейронів-сусідів з решітки мережі адаптуються за правилом Кохонена.

Після того, як вхідний набір даних поданий мережі кілька разів і мережу пройшла навчання на цьому наборі вагові вектори стають розташованими навколо інших вагових векторів, що утворюють центри кластерів. За рахунок того, що мережа спочатку являє собою двовимірну грати її нейрони і значення їх вагових векторів можна представити у вигляді спеціальних карт-розмальовок, що містять кольорові області - карт Кохонена.

Переваги:

- автоматична кластеризация даних;
- візуалізація багатовимірних даних на площині з двома осями;

- навчання без вчителя.

Недоліки:

- складність визначення метрики, за допомогою якої оцінюється відстань між векторами даних.

Мережа з нечіткою самоорганізацією к-середніх. Алгоритм працює аналогічно чіткому алгоритму к-середніх. Відмінність в тому, що кожен вектор має ступінь приналежності до кластерів, а не належить повністю тільки одному кластеру. Таким чином, вектори на кордонах кластерів можуть належати кластеру в меншій мірі, ніж вектори, близькі до центру кластера.

Переваги:

- швидкість процесу навчання;
- гарантована збіжність рішення до глобального мінімуму.

Недоліки:

- кількість кластерів об'єктів встановлюється заздалегідь як константа;
- зростає складність реалізації за рахунок застосування апараті нечіткої логіки.

Асоціативні правила. Пошук асоціативних правил, що генеруються на основі всіх наявних різноманітних наборів комбінацій ознак. Аналізовані дані апріорно розбиваються на набори різних комбінацій ознак. При цьому одні й ті ж ознаки можуть міститися в різних наборах. Далі обчислюється частота тієї, що зустрічається кожної ознаки окремо в усіх наборах. Потім задається мінімальний рівень підтримки. якщо частота тієї, що зустрічається якої-небудь ознаки нижче встановленого рівня підтримки, то він виключається з подальшого аналізу.

На наступному кроці обчислюється частота народження кожної комбінації з двох ознак у всіх наборах. Потім також виключаються з подальшого розгляду ті комбінації, частота яких нижче рівня підтримки. На наступному кроці обчислюється частота народження кожної комбінації з трьох ознаки т.д. Алгоритм зупиняється, коли комбінації ознак перестануть зустрічатися в наборах. Потім виконується генерація правил про те, які

комбінації ознак можуть зустрічатися в різних наборах.

Переваги:

- онлайн обробка інформації.

Недоліки:

- необхідно використовувати базу даних для зберігання різних наборів ознак, інакше простим перебором при великому кількості наборів завдання може бути нерозв'язною;
- не враховується тимчасової аспект появи наборів ознак, а тільки факт настання події.

Послідовні шаблони. Забезпечують розширення можливостей застосування асоціативних правил з урахуванням тимчасового аспекту, послідовності появи ознак в наборах.

Переваги:

- онлайн обробка інформації;
- чи враховується тимчасової аспект виникнення наборів ознак.

Недоліки:

- необхідна наявність бази даних для зберігання наборів ознак.

Тимчасові ряди. Прогнозування майбутніх значень часового ряду за поточними або попереднім значенням. Часовий ряд представляє собою послідовність спостережень за змінами в часі значень ознак (атрибутів) будь-якого об'єкта або процесу. Для аналізу часових рядів використовуються ймовірністно-статистичні моделі, при цьому часовий ряд розглядається як випадковий процес, основними характеристиками якого є математичне очікування, дисперсія, автокореляційна функція часового ряду.

Лінійна і нелінійна регресія. Пошук взаємозв'язку між вхідними і вихідний змінними на основі рівняння регресії.

Для лінійної регресії (при n = 1) коефіцієнти регресії обчислюються за допомогою методу найменших квадратів, а гіперплоскість перетворюється в лінію.

Для нелінійної регресії коефіцієнти регресії обчислюються за

допомогою методів градієнтного спуску.

Переваги:

- швидкість і простота реалізації.

Недоліки:

- добре підходять для опису багатьох процесів тільки в областях економіки і фінансів.

При використанні класичної моделі СОКК виникає ряд проблем, що згадуються в джерелах, до числа яких відносяться "Граничний ефект" і наявність "мертвих" нейронів. Ці проблеми не стільки обмежують можливості застосування нейронних мереж Кохонена, скільки впливають на точність і якість їх роботи.

Спроби удосконалити самоорганізовані карти впливом на їх топологію можна знайти в ряді робіт, в яких пропонується підхід, згідно з яким розмір або структуру мережі СОКК можна зробити залежними від деяких проміжних результатів (наприклад, від помилки квантування в ході процесу навчання мережі). наприклад, кількість вузлів (нейронів) буде подвоюватися після закінчення процесу настройки, якщо в результату не досягнуто необхідна точність. Нові вузли поміщаються в проміжки між вже налаштованими вузлами і мережу донастроювати. Іншим прикладом є роботи, які на додаток до одновимірних і двовимірних карт розглядають також решітки будь-якої розмірності.

На сьогоднішній день по моделі СОКК і її застосуванням у світі опубліковано кілька тисяч робіт. Модель отримала широке застосування в найрізноманітніших областях і стала використовуватися як інструмент для обробки даних різної природи. При цьому області її застосування продовжують розширюватися. найбільш поширеними завданнями для СОКК є завдання класифікації і кластеризації, візуалізації багатовимірних даних, нелінійного проектування, аналізу часових рядів, обробки зображень, розпізнавання мови, діагностики в індустрії і медицині та ін.

1.2 Класична модель СОКК і алгоритм її навчання

Багато частин мозкових тканин, наприклад кора головного мозку, просторово організовані таким чином, що виявляється чітка геометрична локалізація нервових функцій і їх зв'язку зі значеннями ознак, характерних для даної області мозку. біологічна організація нейронів головного мозку має шарувату структуру, яка задається генетично. Сигнали, ініційовані подразниками зі схожими ознаками, які надходять в таку нейронну структуру, збуджують в ній нейрони, близько розташовані один до одного. результатом роздратування нейрона служить певний електричне збудження нервової клітини, яке, в свою чергу, може стати подразником для іншого нейрона. Внаслідок такої взаємодії між нейронами в головному мозку формується топографічне відображення ознак подразників, яке відношенню близькості вхідних ознак ставить у відповідність просторову близькість реагують нейронів.

Намагаючись реалізувати принцип навчання, який був би дійсно працездатним при вирішенні практичних завдань та може ефективно формувати глобально впорядковані відображення різних ознак на шарувату нейронну мережу, професор обчислювальної техніки, професор Академії наук Фінляндії Тейво Кохонен в 1981-1982 рр. формалізував процес самоорганізації у вигляді алгоритму, який називається самоорганізаванними картами Кохонена.

СОКК – це нейронна мережа без зворотніх зв'язків, в якій використовується алгоритм навчання без вчителя. За допомогою процесу, іменованого самоорганізацією, СОКК утворює топологічне представлення вхідних даних, що аналізуються з нейронів, одержуваних на виході. СОКК можна навчити дізнаватися або знаходити взаємозв'язки між входами і виходами або організовувати дані таким чином, щоб виявляти в них раніше невідомі образи або структури.

Метод навчання СОКК не припускав зовнішнього втручання. В

нейромережевих методиках, які передбачають навчання з учителем, для знаходження образу або співвідношення між вхідними даними потрібно, щоб один або більше виходів були точно задані разом з одним або більше входами. СОКК, навпаки, відображає дані більшої розмірності на карті меншої розмірності, що складається з решітки нейронів.

Алгоритм самоорганізації Кохонена грунтується на змагальному навчанні без учителя. Він забезпечує зберігає топологію відображення з простору великої розмірності в нейрони карти, які зазвичай утворюють двовимірну решітку. Таким чином, це відображення є відображенням простору великої розмірності на площину. Властивість збереження топології означає, що СОКК розподіляє подібні вектори вхідних даних по нейронам, тобто точки, розплоджені в просторі входів близько один до одного, відображаються на близько розплоджені нейрони мережі. Таким чином, СОКК може служити як засобом кластеризації, так і засобом візуального представлення даних великої розмірності.

У своєму класичному варіанті СОКК має всього два прошарки: вхідний шар, що містить нейрони для кожного вектора вхідного простору аналізованих даних, і вихідний шар нейронів, пов'язаних з усіма нейронами вхідного шару за допомогою вагових векторів. Вихідний шар називається також шаром топографічної карти. нейрони топографічної карти розташовуються, як правило, в двовимірному просторі. Число нейронів в топографічній карті визначається користувачем на підставі початкової форми або розміру карти, яку він хоче отримати. Нейрони в карті пов'язані так званими латеральними зв'язками (По аналогії з біологічними зв'язками між нейронами в головному мозку). Чим далі нейрони розташовані один від одного, тим менше цей зв'язок.

Коли вхідний вектор подається мережі, нейрони вихідного шару змагаються один з одним за право бути переможцем. переможцем стає той вихідний нейрон, ваги зв'язків якого виявляються найближчими до вхідного образу в сенсі евклідова відстані. Після того як вхідний вектор пред'явлений

мережі, кожен нейрон прагне досягти найбільшого з ним відповідності. Вихідний нейрон, найближчий до вхідного образу, визнається переможцем. Вагові вектори нейрона-переможця потім коригуються, тобто зсуваються в напрямку вхідного вектора за допомогою множника, який визначається коефіцієнтом навчання. В цьому і полягає сутність змагальних нейронних мереж.

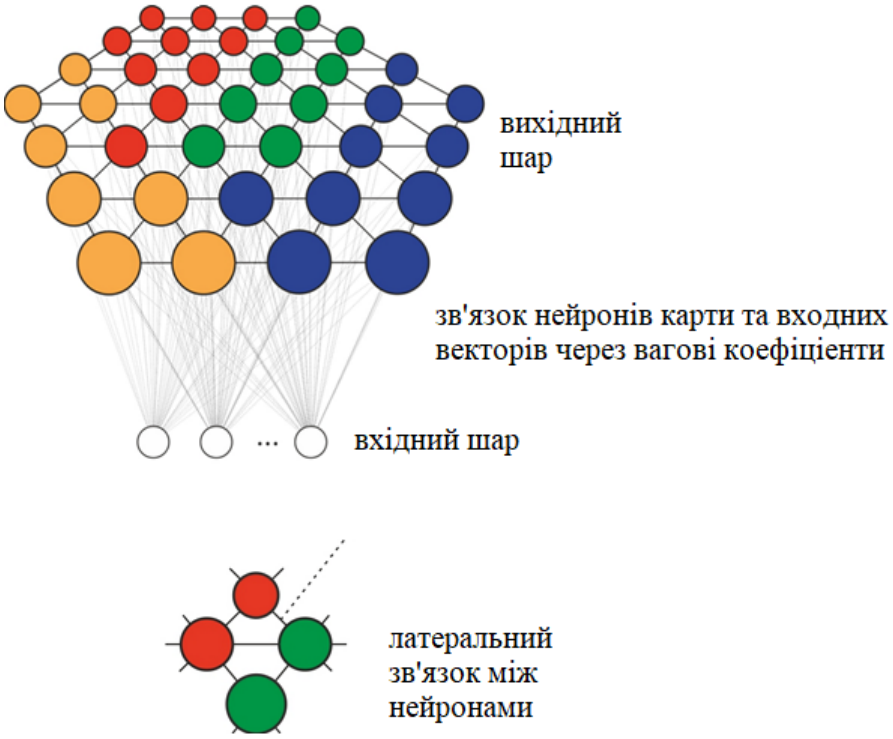


Рисунок 1.1 – Два шари мережі Кахонена класичної моделі

Коли СОКК здійснює топологічне відображення, відбувається регулювання не тільки ваги нейрона-переможця, але також ваг суміжних вихідних нейронів, найближчих сусідів переможця. ціла околиця вихідних нейронів стає зрушеною ближче до вхідного образу. Коли процес починається з випадкових значень ваг, вихідні нейрони повільно вирівнюються, оскільки при пред'явленні вхідного образу на нього реагує не тільки окремий нейрон, але також і його околиці. У міру того як навчання триває, розмір цієї околиці навколо нейрона-переможця поступово

зменшується. В кінці навчання коригуються тільки ваги нейрона-переможця.

Результатом є ваги зв'язків між вхідними векторами і вихідними нейронами, кожному з яких відповідає типовий вхідний образ для деякого підмножини вхідних даних, яке потрапляє в окремий кластер. Процес стиснення даних великої розмірності до деякого набору кластерів називається сегментацією. початкове простір великої розмірності стискається в двовимірну карту. Індекс вихідного нейрона-переможця, по суті, розділяє вхідні образи на безліч категорій або кластерів.

СОКК також мають здатність до узагальнення. Це означає, що подібні нейронні мережі можуть дізнаватися або характеризувати вхідні дані, з якими вони ніколи раніше не мали справу. новий вектор вхідних даних співвідноситься з тим елементом карти, на який він відображається. Більш того, для пошуку або прогнозування значень пропущених даних на основі використання раніше навченої карти вони можуть використовувати навіть вхідні вектори з відсутніми (пропущеними) даними.

Ітераційний алгоритм навчання СОКК. Вихідний ітераційний алгоритм СОКК визначає рекурсивний процес спеціального виду, в якому на кожному кроці здійснюється обробка тільки частини нейронів.

Нехай в евклідовому просторі R^m з системою координат $(x^1,...,x^m)$, де m - розмірність простору, задана фізична область X, на якій визначено набір вхідних векторів, який позначається $X_K = \{x_1, x_2, ..., x_K\}$, де К – кількість вхідних векторів, $x_i \in X$, $x_i = (x_i^1,...,x_i^m)$, i=1,…,K. Нехай U – це обчислювальна область в пропросторі R^2 з системою координат (u^1,u^2), на якій задана мережа вузлів (нейронів) $U_N = \{u_1,...,u_N\}$, де N – кількість нейронів мережі, $u_j \in U$, $u_j = (u_j^1,u_j^2)$, j = 1,...,N. Нейрони в мережі пов'язані між собою зв'язком, який називається латеральним.

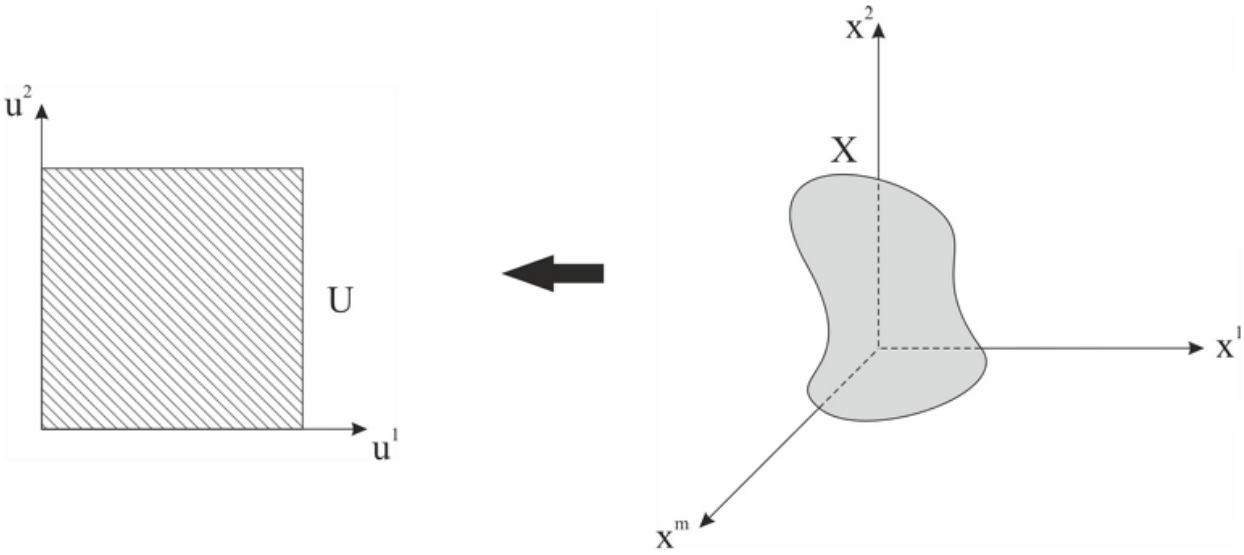


Рисунок 1.2 – Алгоритм СОКК: обчислювальна область U із зафіксовано сіткою U_N,фізична область X з простором вхідних даних X_K.

Алгоритм СОКК визначає відображення простору $X \in R^m$ на $U \in R^2$, яке переводить набір вхідних векторів Хк на двовимірні грати нейронів U_N (рисунок 1.2).

Кожному нейрону u_j ставиться у відповідність параметричний вектор, званий ваговим вектором $w_j = (w_j^1,...,w_j^m)$, де $w_j \in R^m$, j = l,..,,N. На стадії ініціалізації мережі всім ваговим векторах присвоюються невеликі випадкові значення.

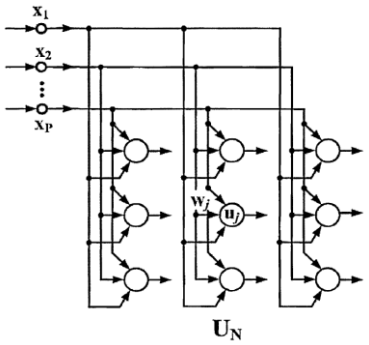


Рисунок 1.3 – Масив вузлів (нейронів) у двомірних гратах U_N СОКК

Для навчання СОКК з Хк вибирають підмножину $P = \{x_1, x_2, ..., x_p\} \in X_K$ навчальних векторів, які, також можуть збігатися з Хк. За допомогою вагового вектора wj – кожен нейрон з'єднаний з усіма вхідними векторами з Р (рисунок 1.3).

При цьому вхідні навчальні вектори x_p нормалізують. При нормалізованих вхідних векторах прагнуть до них вектори ваг wj нормалізуються автоматично.

Після того, як підготовлена структура нейронної мережі, починається процес переміщення нейронів у вхідному просторі R^m за наступним алгоритмом.

а) з навчальної множини Р вхідних даних випадковим чином вибирається елемент, який буде вхідним вектором х для СОКК;

б) у нейронній мережі вибирається нейрон, який переміг в конкурентній боротьбі завдяки тому, що його вектор ваг в найменшій мірі відрізняється від відповідних компонентів вектора х. Для такого w-го нейрона-переможця має виконуватися відношення:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (1.1)$$

Такий спосіб визначення нейрона-переможця є алгоритмом, який називається "переможець забирає все" (англ.: Winner Takes All, WTA).

Найчастіше в якості міри відстані використовується евклідова міра:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (1.2)$$

Іншими заходами відстані є:

- скалярний добуток:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.3}$$

- міра щодо норми L1 (Манхеттен):

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.4}$$

- міра щодо норми Lx:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.5}$$

У даній роботі використана евклідова міра (2) вимірювання відстані між векторами.

Навколо нейрона-переможця утворюється топологічна область сусідства

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.6}$$

Вона складається з нейронів-сусідів (рисунок 1.4). Розміри цієї області з плином часу навчання зменшуються, тобто з кожної итерацией навчання в цю область потрапляє все менша кількість нейронів-сусідів.

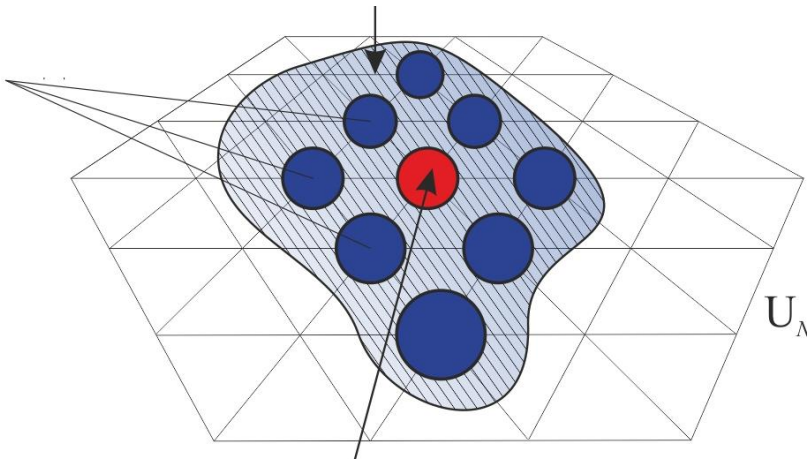


Рисунок 1.4 – Топологічна область сусідства для нейрона-переможця

Нейрон-переможець і все нейрони, що лежать в межах його Sw (t) області, піддаються адаптації, в ході якої їх вагові вектори змінюються в напрямку вектора x за правилом Кохонена:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.7}$$

Як показано вище (формула 1.7) параметр $\eta_i(t)$ позначає коефіцієнт навчання ui - го нейрона з околиці Sw(t) в t-й момент часу. При цьому

Епоха навчання – це один прохід нейронної мережі по всьому навчальному безлічі P вхідних векторів. Кількість епох навчання задається за допомогою параметра T перед початком процесу навчання.

Коефіцієнт навчання $\eta_i(t)$ – це функція, яка приймає значення з проміжку [0; 1) і зменшується зі збільшенням відстані між u_i -м нейроном-сусідом і u_w -м нейроном переможцем в просторі решітки мережі UN e R. Зазвичай вона задається у вигляді добутку двох функцій:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.8}$$

де $\alpha(t)$ – функції швидкості навчання, $h_{wi}(t)$ – функції сусідства.

Функція швидкості навчання $\alpha(t)$ встановлюється однаковою для всіх нейронів і монотонно убуває з часом навчання t мережі. Вона може бути лінійної, експоненціальної або обернено пропорційній часу t. Ефективний вибір цієї функції і її параметрів виконується при роботах з СОКК поки здебільшого експериментально. У цій атестаційній роботі використовується коефіцієнт середньої оптимальної швидкості, який на кожній ітерації навчання обчислюється за формулою:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.9}$$

де $\alpha_0 \leq 1$ – початкова швидкість, що задається користувачем перед навчанням мережі. Автором пропонується задавати значення початкової швидкості навчання, рівне $\alpha_0 = 0,75$.

Для функції сусідства в класичному алгоритмі Кохонена має місце наступне співвідношення, яке називається сусідством прямокутного типу:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.10}$$

де $d(u_w, u_i)$ – евклідова відстань між векторами ваг u_w-го нейрона-переможця та u_i-го нейрона мережі, $\sigma(t)$ – радіус сусідства – ще одна лінійна монотонно спадна функція часу, що задає розмір топологічної околиці $S_w(t)$:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.11}$$

Інший тип сусідства, часто вживаний в картах Кохонена та використовуваний в роботі, – це сусідство гауссовського типу, при якому функція $h_{wi}(t)$ визначається формулою:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.12}$$

Функція сусідства задає вагу латеральної зв'язку між нейронами мережі і грає найважливішу роль в коефіцієнті навчання мережі, задаючи якісні характеристики навченою мережі. фізичний сенс латеральних зв'язків полягає в тому, що при навчанні величина зсуву ваги $w_i(t)$ залежить від латеральної зв'язку між нейроном і (і нейроном-переможцем u_w на ітерації t.

Для збіжності процесу навчання Кохонена необхідно, щоб виконувалася умова $h_{wi}(t) \to 0$ при $t \to \infty$. Для обох функцій сусідства прямокутного і гаусовського типів зі зростанням $d(u_{w,}u_i)$ виконується умова $h_{wi}(t) \to 0$. Середня ширина і форма функції сусідства визначають "Жорсткість" тієї "еластичною поверхні", яка підганяється так, щоб найкращим чином відповідати векторах вхідних аналізованих даних. Показано, що метастабільні стани, що представляють топологічні дефекти в конфігурації навченої мережі, виникають в тих випадках, коли алгоритм СОКК використовує не опуклого функцію околиці. Функція Гауса є опуклою, в той час як прямокутна функція – ні. Широка, опукла функція околиці (така як Функція Гауса великого радіусу) призводить до більш швидкого топологічного упорядкування.

Область сусідства S_w нейрона-переможця u_w визначається як безліч нейронів, що є сусідами з нейроном-переможцем в решіті:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.13}$$

На рисунку 1.5 показано кілька прикладів областей сусідства різних радіусів.

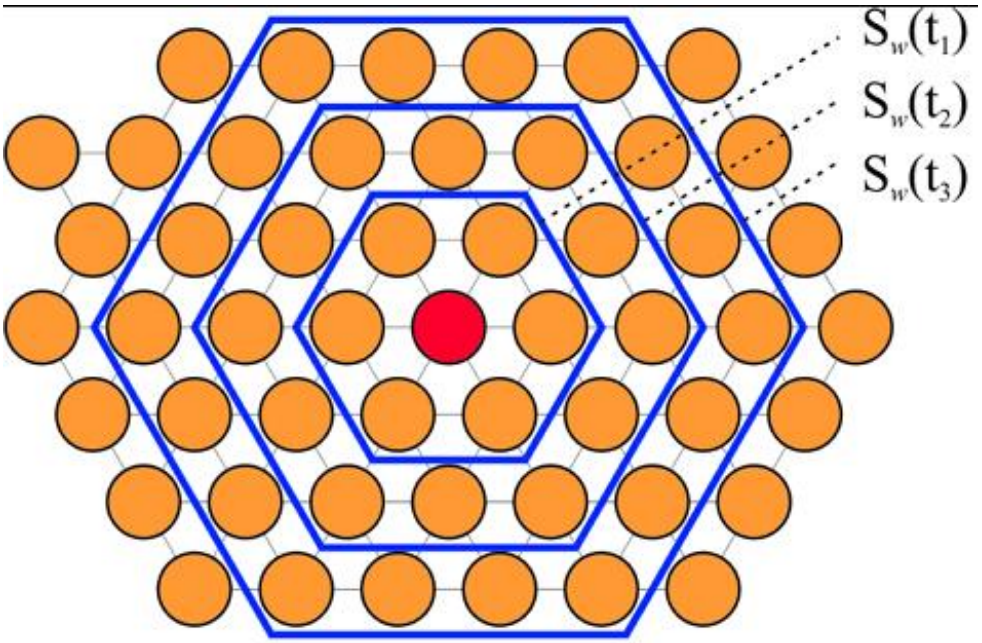


Рисунок 1.5 – Області сусісдства нейрона-переможця різних радіусів

Алгоритм навчання СОКК повторюється, повертаючись до вибору наступного випадкового вхідного вектора з навчальної множини P. Кількість повторень алгоритму дорівнює кількості епох навчання T.

Оскільки навчання є випадковий процес, кінцева статистична точність відображення залежить від кількості епох на завершальному етапі даного процесу, який повинен бути досить тривалим, причому обійти цю вимогу можливості немає. Евристичне правило отримання хорошої статистичної точності полягає в тому, що кількість ітерацій процесу навчання має, принаймні, в 500 разів перевищувати кількість нейронів мережі.

Навчання навмисне розбивають на дві фази: більш коротку – з великою швидкістю навчання a й великим радіусом навчання a, і більш довгу - з малою швидкістю навчання і майже нульовими радіусом сусідства.

Після закінчення T епох процес навчання СОКК зупиняється і виконується оцінка похибки її навчання, тобто похибки апроксимації

вхідного простору, звана помилкою квантування :

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (1.14)$$

При цьому кожен вектор $x_i (i = 1,...,p)$ з безлічі P подається по черзі вже навченої нейронної мережі, для нього відшукується w-ий нейрон-переможець і обчислюється квадрат відстані між ними. Середньозважена сума квадратів таких відстаней для всіх векторів і дає помилку квантування. Вибір певного рівня значущості помилки квантування, вище якого результати навчання СОКК відкидаються як помилкові, є досить довільним. На практиці остаточне рішення зазвичай залежить від того, чи був результат передбачений апріорі або виявлений апостериорно в результаті навчання СОКК. У даній роботі прийнятними значеннями помилки квантування вважаються значення менші або рівні 0,05 (Eq <0,05).

В процесі навчання СОКК вагові вектори нейронів налаштовуються таким чином, щоб нейрони розташовувалися в місцях локальних згущень вхідних даних, тобто описували кластерну структуру хмари даних. В цей же час зв'язку між нейронами відповідають відносинам сусідства між відповідними кластерами у вхідному просторі.

Після навчання нейронна мережа здатна класифікувати поданий їй на вхід випробуваний вектор, знаходячи найближчий до нього нейрон-переможець і, тим самим, визначаючи кластер, до якого цей нейрон-переможець належить. Всі кластери нейронної мережі маркуються змістовними за змістом мітками. Якщо ж кластер знайти не вдається, то це означає, що мережа не прийняла ніякого рішення і подається вхідний вектор належить новому які раніше не пізнаному кластеру. Таким чином, відбувається пошук нових знань в аналізованих даних.

Після того, як мережа навчена розпізнаванню структури даних, її можна використовувати як засіб візуалізації при аналізі даних. Можна

обробляти окремі спостереження і виявляти, як при цьому змінюється топологічна карта. Це дозволяє зрозуміти, чи мають кластери якийсь змістовний сенс (як правило, при цьому доводиться повертатися до змістовного змістом завдання, щоб встановити, як співвідносяться один з одним кластери аналізованих даних).

1.3 Підготовка і попередня обробка даних для нейронної мережі

Для вирішення завдань за допомогою нейронних мереж необхідна попередня обробка вхідних даних, які будуть надані їй для навчання.

Основними етапами попередньої обробки даних є:

- кодування вхідних векторів. Нейронні мережі можуть працювати тільки з числовими даними;
- нормування даних. Результати нейромережевого аналізу не повинні залежати від вибору одиниць виміру;
- передобробка даних. Видалення очевидних регулярностей з даних полегшує нейромережі виявлення нетривіальних закономірностей.

Кодування вхідних векторів. Виділяються два основних типи нечислових змінних: впорядковані і категоріальні. В обох випадках змінна відноситься до одного з дискретного набору класів $\{c_i,...,c_n\}$. Для впорядкованих змінних і класи ці впорядковані - їх можна ранжувати: $c_1 > c_2 > ... > c_n$.Для категоріальних змінних така впорядкованість відсутня. Прикладами упорядкованих змінних можуть бути порівняльні категорії: погано-добре-відмінно, повільно-швидко. Категоріальні змінні позначають один з класів, будучи "іменами" категорій. Це можуть бути імена людей або назви кольорів: білий, синій, червоний.

Впорядковані змінні ближчі до числової формі, так як числовий ряд вже впорядкований. Для кодування таких змінних необхідно поставити у відповідність номерам категорій такі числові значення, які зберігали б існуючу впорядкованість.

Кодування категоріальних змінних також можна нумерувати довільним чином. Описані методи двійкового кодування і категоріальних змінних.

Нормування даних Вектори вхідних даних можуть мати різний масштаб. Приведення даних до одиничного масштабу забезпечується нормуванням кожної компоненти $x^i \mid i = 1,...,m$ вхідного вектора на діапазон розкиду її значень, до векторах вхідного набору даних. В найпростішому варіанті це – лінійне перетворення:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.15}$$

Лінійне нормування оптимальне, коли значення і -ої компоненти всіх векторів вхідного набору даних щільно заповнюють певний інтервал. Але якщо в даних є відносно рідкісні викиди, набагато перевищують типовий розкид.

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.16}$$

де

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.17}$$

та

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.18}$$

Однак, в цьому випадку (формула 1.16), нормовані величини не

належать гарантовано одиничного інтервалу:

Для гарантованої нормування значень компоненти x1 вхідних векторів по інтервалу [0,1] пропонується до використання наступне нелінійне перетворення:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (1.19)$$

У даній роботі виконується нормування вхідних векторів даних, використовуючи перетворення (19), що забезпечує нормування за шкалою [0,1].

Попередня обробка даних. Сильною стороною нейромережевого аналізу є можливість отримання передбачень при мінімумі апріорних знань. Оскільки заздалегідь зазвичай невідомо наскільки корисні ті чи інші змінні (Компоненти), що описують вхідні вектори, у дослідника виникає спокуса збільшувати число вхідних параметрів, в надії на те, що мережа сама визначить які з них найбільш значущі. Але зі збільшенням розмірності вхідного вектора відбувається зменшення точності прогнозів.

Один з найбільш простих і поширених способів зниження розмірності - використання головних компонент вхідних векторів.

Цей метод дозволяє не відкидаючи конкретні компоненти враховувати лише найбільш значущі комбінації їх значень, переходячи до нового ортогональному базису, осі якого орієнтовані за напрямками максимальної дисперсії набору вхідних даних.

Уздовж першої осі нового базису дисперсія максимальна, друга вісь максимізує дисперсію при умови ортогональності першої осі, і т.д., остання вісь має мінімальну дисперсію з усіх можливих.

Таке перетворення дозволяє знижувати інформацію шляхом відкидання координат, відповідних напрямками з мінімальною дисперсією.

Зниження розмірності входів за допомогою застосування нейронних

мереж, а також відновлення пропущених компонент даних за допомогою методу головних компонент також описуються в джерелах.

1.4 Побудова візуальних топографічних карт для СОКК

Алгоритм самоорганізації Кохонена забезпечує зберігаюче топологію відображення з простору великої розмірності в нейрони мережі, які зазвичай утворюють двовимірну решітку. Таким чином, це відображення є відображенням простору великої розмірності на площину.

Властивість збереження топології означає, що СОКК розподіляє подібні вектори вхідних даних по нейронам, тобто точки, розплоджені в просторі входів близько один до одного, відображаються на близько розплоджені нейрони мережі. Завдяки цьому, СОКК може служити як засобом кластеризації, так і засобом візуального представлення даних великої розмірності.

Нижче перераховані основні способи побудови візуальних топографічних карт навчених самоорганізованих нейронних мереж Кохонена.

Розфарбовування ділянок карти в різні кольори за допомогою кольорової палітри або відтінків сірого кольору.

При цьому колір кожної ділянки відображає середню відстань між ваговим вектором нейрона, яким відповідає розфарбовувати ділянку карти, і векторами ваг найближчих сусідів цього нейрона.

Таке відображення називається U-матрицею.

Якщо середня відстань між сусідніми ваговими векторами мало, то використовуються, наприклад, сині відтінки кольору, і навпаки, великі відстані між векторами відображаються відтінками червоного кольору.

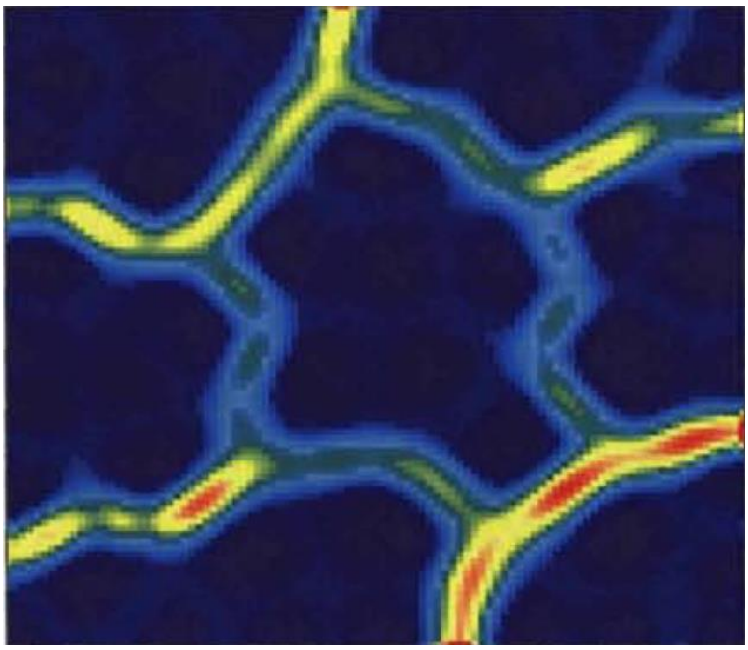


Рисунок 1.6 – Карта СОКК, побудована на технології U-матриці

Ділянки карт також можуть маркуватися. Якщо у вхідному наборі даних є вектори з присвоєними їм мітками, то ці мітки ставляться на ділянку карти, нейрон якого знаходиться найближче до такого вхідного вектора (рисунок 1.7).

1	1	1	1	2	2	2	3	3	4
1	1	1	2	2	2	3	3	3	4
1	1	2	2	2	2	3	3	3	4
1	1	2	7	7	3	3	3	4	4
8	8	8	7	7	7	3	4	4	4
8	8	7	7	5	7	3	4	4	4
8	6	7	7	5	4	4	4	4	4
8	6	6	7	5	5	5	4	4	4
6	6	6	6	5	5	5	5	5	5
6	6	6	6	6	6	5	5	5	5

Рисунок 1.7 – Карта СОКК. Ділянках карти присвоєні маркери вхідних векторів, яким вони відповідають

Розфарбовування ділянок карт за значенням якої-небудь ознаки або компоненти вагових векторів (рисунок 1.8). При цьому колір кожної ділянки відображає значення цієї ознаки або компоненти. Низьким значеннями ознаки відповідають сині кольори, а високих значень – червоні кольори. При цьому під картою відображається кольорова градієнтна шкала, на якій проставлені числові значення для кожного кольору карти.

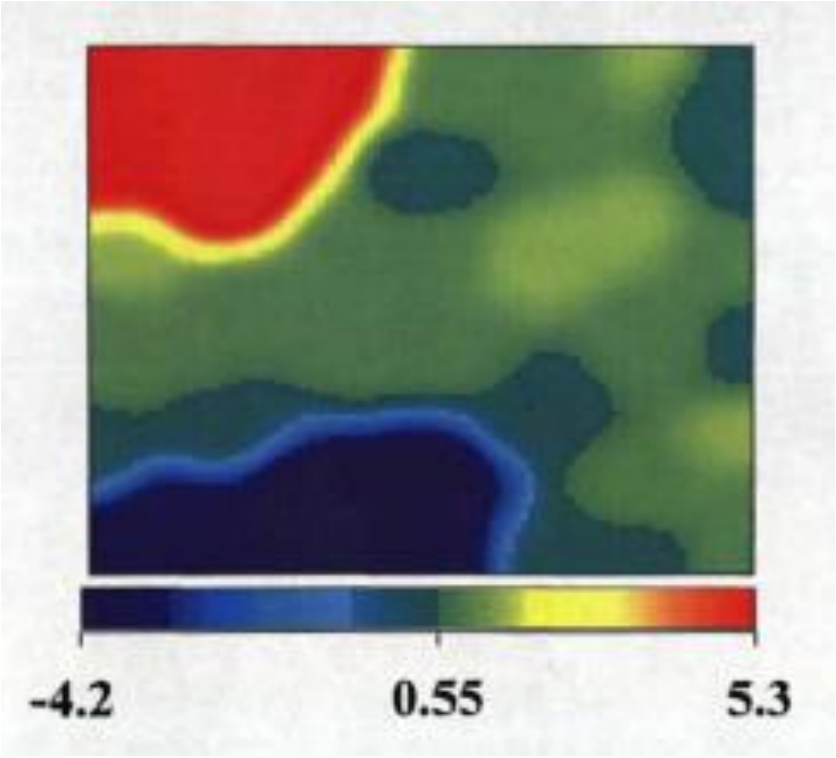


Рисунок 1.8 – Карта СОКК, побудована за значенням якої-небудь ознаки або компоненти

Відображення на кожній ділянці карти квадрата, розмір якого пропорційний числу точок даних, найближчих з точки зору вагових векторів до нейрона, який відображається на даній ділянці. Квадрат розфарбовується в колір, який відповідає значенню відображається компоненти вагового вектора. Такий спосіб відображення називається діаграмою Хінтона (рисунок 1.9).

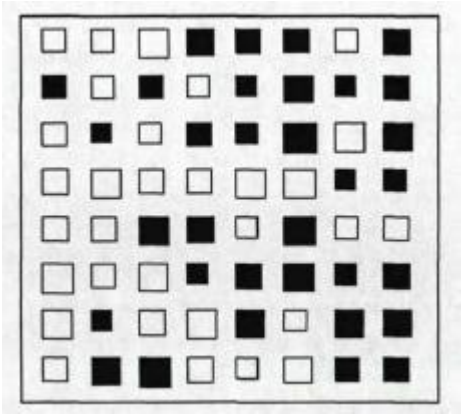


Рисунок 1.9 – Карта СОКК, побудована як діаграма Хінтона

Слід ще раз підкреслити, що СОКК виконує топологічний відображення багатовимірного простору вхідних даних на плоску карту. Це означає, що близьким елементів вхідного простору відповідає один і той же або близькі нейрони на карті.

Таким чином, вивчаючи розмальовки карт компонент і кластерів, можна виявляти закономірності, якими володіють аналізовані дані, а також класи (кластери), на які ці дані поділяються. Більш того, отримавши набір карт для декількох серій аналізованих даних, виконується аналіз тенденцій пересування даних між кластерами.

Розфарбування карт компонент вагових векторів нейронів. За рахунок мульті розмірності вагових векторів нейронів СОКК можна будувати карти, кожна з яких буде відображати характер розподілу даних в розрізі окремої компоненти одночасно всіх цих векторів (рисунок 1.10).

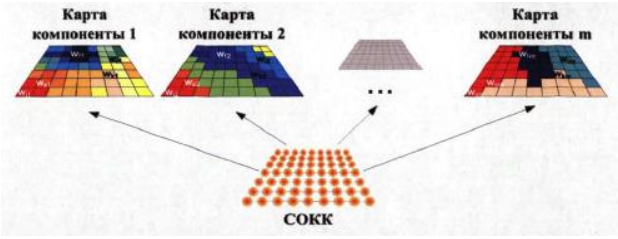


Рисунок 1.10 – Карти компонент СОКК

Для визначення кольору осередків сітки карти кожної компоненти виконуються наступні дії.

Створюється масив розміру R градієнта кольорів, кожен елемент якого є кольором. Кольори в масиві змінюються відповідно до індексом масиву. Значенням індексів масиву від 1 до R відповідають кольори від фіолетового до червоного відповідно.

Вибирають компоненту вагових векторів всіх нейронів СОКК, для якої відображається карта.

Числові значення компоненти для всіх нейронів мережі упорядковано по зростанню і масштабуються за шкалою від 1 до R колірного градієнта.

Кожний осередок карти компоненти, відповідного нейрона СОКК, розфарбовується в колір, який відповідає числовому значенню цієї компоненти відображуваного нейрона, відповідно до масивом кольорів (рисунок 1.11).

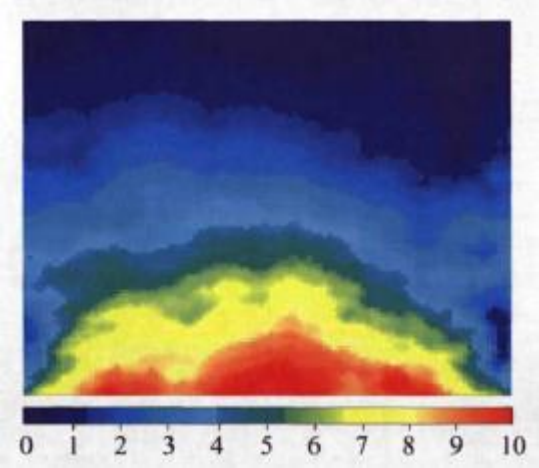


Рисунок 1.11 – Карта однієї компоненти вагових векторів всіх нейроні мережі (Розмір мережі 100x100 нейронів)

Розфарбування карти кластерів СОКК проводиться за технологією U-матриці. Відстань між ваговим вектором нейрона, для якого відображається поточна комірка карти і ваговими векторами його нейронів- сусідів

розраховується за формулою:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{1.20}$$

Нейрони-сусіди можуть вибиратися з околиці сусідства будь-якого радіусу. В роботі використовується радіус такої околиці, рівний 1 (рисунок 1.13).

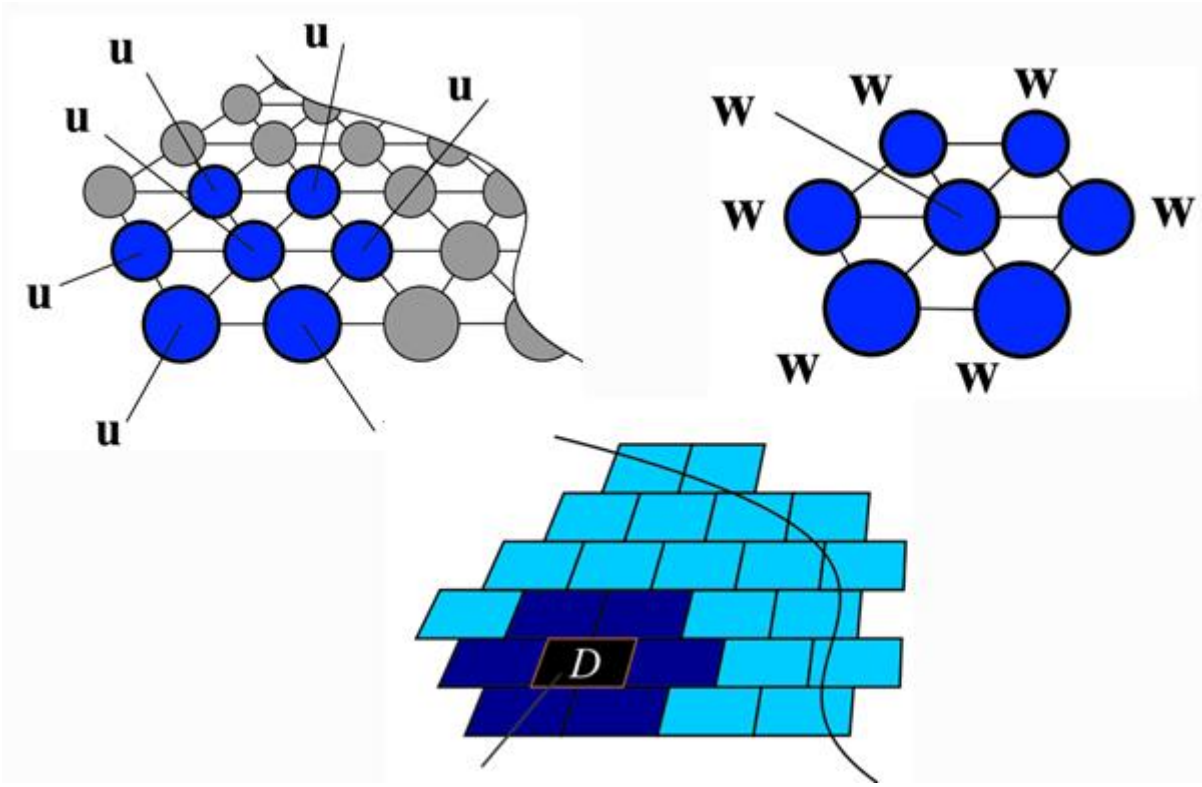


Рисунок 1.12 – Розмальовка ділянки карти кластерів СОКК

Класична модель модель самоорганізованої нейронної мережі, запропонована Кохоненом, надає можливості кластерного аналізу багатовимірних даних і є засобом візуального відображення характеру їх розподілу на двовимірної топологічної карті.

На рисунку 1.13 зображені проблеми застосування класичних мереж Кохонена.

Рисунок 1.13 - Проблеми застосування класичної моделі карти Кохонена

Класична модель самоорганізованої нейронної мережі, запропонована Кохоненом, надає можливості кластерного аналізу багатовимірних даних і є засобом візуального відображення характеру їх розподілу на двовимірній топологічній карті.

2 МЕТОД ВДОСКОНАЛЕННЯ АПАРАТУ КАРТ КОХОНЕНА

У цьому розділі наводиться аналіз впливу відомих проблем класичної моделі СОКК на якість її навчання. Пропонуються новий варіант структури решітки мережі і новий спосіб визначення сусідства між нейронами в такий решітці, за допомогою яких можна вирішити ці проблеми. Наводяться результати досліджень нової моделі на основі порівняльного аналізу результатів її навчання і мережі класичної моделі цього виду за двома критеріями: помилки квантування і ентропії. В кінці глави наводяться результати порівняльного аналізу за цими ж критеріям між розробленої мережею і її існуючим аналогом.

2.1 Проблеми застосування класичної моделі СОКК

Проблема мертвих нейронів. При ініціалізації ваг мережі випадковим способом частина нейронів може виявитися в області простору, в якій відсутні дані або їх кількість мізерно мало. Ці нейрони мають мало шансів на перемогу і адаптацію своїх ваг, тому вони залишаються мертвими. Таким чином, вхідні дані будуть інтерпретуватися меншою кількістю нейронів (мертві нейрони не беруть участі в аналізі), а похибка інтерпретації даних, яку називають похибкою квантування, збільшиться. Тому важливою проблемою стає активація всіх нейронів мережі.

Таку активацію можна здійснити, якщо в алгоритмі навчання передбачити облік кількості перемог кожного нейрона, а процес навчання організувати так, щоб дати шанс перемогти і менш активним нейронам. Такий спосіб обліку активності нейронів називається механізмом стомлення.

У даній роботі використаний метод підрахунку потенціалу pt кожного нейрона, значення якого модифікується щоразу після уявлення черговий реалізації вхідного векторах відповідно до наступною формулою (в ній

передбачається, що переможцем став w-й нейрон):

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.1)$$

Значення коефіцієнта p_{min} визначає мінімальний потенціал, що дозволяє участь в конкурентній боротьбі. Якщо фактичне значення потенціалу p_i падає нижче p_{min}, і-й нейрон "відпочиває", а переможець шукається серед нейронів, для яких виконується відношення для $1 \le i \le N$ та $p_i \ge p_{min}$. Максимальне значення потенціалу обмежується на рівні, рівному 1. На практиці хороші результати досягаються, коли p_{min} та 0,75. Це значення мінімального потенціалу і використовується при навчанні мереж в даній роботі.

Граничний ефект. У процесі навчання всі нейрони мережі змагаються один з одним за вхідні дані; нейрон-переможець і його сусіди адаптують свої ваги. В ідеалі, в результаті навчання все вхідні вектори в рівній мірі представлені нейронам, і сусідні області в решітці мережі мають тенденцію моделювати схожі області вхідного простору. Однак нейрони на краях решітки СОКК мають меншу кількість сусідів, ніж нейрони, що знаходяться всередині карти. Внутрішні нейрони, часто виграючи, захоплюють нейрони з країв мережі в області сусідства, адаптуючи частіше їх вагові вектори в свою сторону. Таким чином, вхідний простір, відображене на мережі, "мнеться" до її центру.

2.2 Способи усунення граничного ефекту

Застосування віртуальних нейронів. Граничний ефект виникає через несиметричності латеральних зв'язків у нейронів, які розташовано близько до кордону обчислювальної області U. Зажадаємо, щоб грати нейронної мережі Un була рівномірною прямокутної, що, взагалі кажучи, необов'язково для

СОКК. Нехай u_i - це внутрішній нейрон, для якого відстань до кордони обчислювальної області більше радіуса навчання σ. Структура мережі Un задана так, що bcs нейрони-сусіди uj з області сусідства Si (13) розташовуються симетрично щодо u_i. Тоді значення ваг латеральних зв'язків h_y (12) також розташовуються симетрично щодо u_i. Якщо при цьому враховувати умову рівноправності нейронів в мережі, то на нейрон w, з однаковою ймовірністю може бути надано вплив будь-яким іншим нейроном u_j. Це означає, що вектор ваги w, цього нейрона в просторі X має однакову ймовірність зрушити симетрично в усіх напрямках під впливом нейронів з Si. Передбачається, що взаємний вплив між нейронами u_i та $u_j \notin S_i$ достатньо малий.

Якщо відстань від u_i до кордону обчислювальної області менше , то в області сусідства Si розташовано недостатньо нейронів для симетрії. У цьому випадку більшість нейронів в Si, - змушують вектор wi , рухатися в основному до центру області X. Тому для балансування асиметрії вектор wi, змушений відсуватися від кордону X.

Для оцінки асиметрії пропонується розглядати наступну характеристику для нейрона u_i:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.2)$$

Для кожного нейрона ця характеристика дорівнює сумі ваг латеральних зв'язків з усіма іншими нейронами з області сусідства S ;. Якщо щ розташований близько до кордону U, то в сумі недостатньо доданків. Тому S, убуває поблизу кордону U.

Для балансування граничного ефекту автор пропонує методику, яка дозволяє використовувати граничні вузли в якості представників відсутніх нейронів за межами обчислювальної області U. Для цього використовуються

віртуальні нейрони, які розташовуються поруч з граничними нейронами мережі, для них задаються спеціальні ваги латеральних зв'язків (12), що залежать від їх положення. Навчання мережі відбувається з урахуванням наявності віртуальних нейронів, які можуть також ставати переможцями і впливають на напрямок зміни вагових векторів граничних нейронів мережі. Підкреслюється, що віртуальні нейрони не існує в алгоритмі, тому фізично структура мережі Un не змінюється. Крім того, зберігається внутрішній паралелізм алгоритму СОКК, які полягає в тому, що всі нейрони обробляються по одному і тому ж правилу незалежно один від одного. В результаті застосування методики величина асиметрії (22) для всіх нейронної мережі стає майже постійною.

Застосування правила зважування. Для усунення граничного ефекту для вхідних векторів низькою розмірності запропоновано правило зважування, коли вхідні вектори постачають умовними вагами, які, в свою чергу, по-різному впливають на коефіцієнт швидкості навчання для нейронів знаходяться на краях мережі і всередині неї. Так в процесі навчання при пред'явленні нейронної мережі чергового вхідного вектора x необхідно здійснити вплив на коефіцієнт швидкості навчання a(t) нейронної мережі (9) наступними вагами:

а). Вагою $W_1 > 1$, коли вектор wi, деякого кутового нейрона ui решітки мережі є найближчим до x і радіус навчання $\sigma(t)$ дорівнює 0, тобто навчання за правилом Кохонена піддається тільки сам нейрон ui без участі його нейронів-сусідів:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.3)$$

б). Вагою $W_2 > 1$, коли вектор wj. деякого нейрона uj знаходиться на кордоні решітки нейронної мережі, але не є кутовим нейроном:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.4}$$

При цьому навчання за правилом Кохонена (7) піддається не тільки нейрон uj, але і його найближчі топологічні нейрони-сусіди, також лежать на кордоні решітки нейронної мережі, включаючи кутові нейрони. Щоб гарантувати стійкість процесу навчання СОКК, потрібно виконання умови

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.5}$$

Тому даний прийом не застосуємо на перших епохах навчання, коли a(t) ще велике. Пропонується використовувати значення $W_1 = 81, W_2 = 9$.

Застосування сферичної топології мережі. В роботі [6] австралійськими авторами пропонується методика усунення проблеми "граничного ефекту" для самоорганізується мережі шляхом трансформації її решітки в замкнуту поверхню, подібну Ікосаедр, кожна грань якого є рівносторонній трикутник (рисунок 2.1).

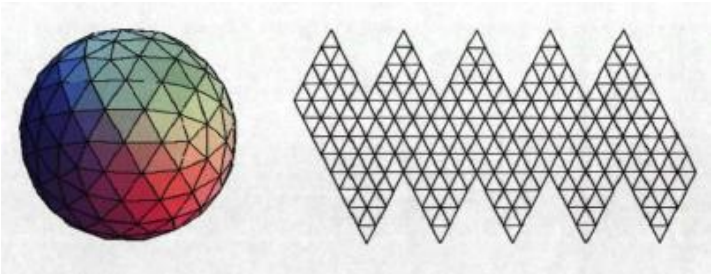


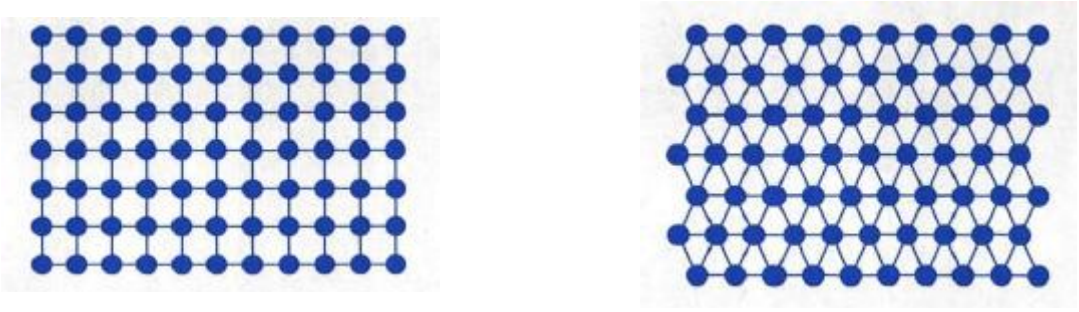
Рисунок 2.1 – Ікосаедр і його розгортка

За рахунок трикутної топології і з'єднання нейронів на краях плоскою решітки СОКК все нейрони мережі мають в заданій області сусідства однакову кількість симетрично оточених нейронами-сусідами.

Такі мережі були названі Geodesic Self-Organizing Map (GEOSOM). Нейрони мережі розташовуються в вершинах трикутників, що утворюють замкнуту поверхню. Для виконання процесу самоорганізації GEOSOM розгортається в плоску решітку, на основі якої відбувається обчислення відстаней між нейронами мережі і виконується адаптація їх вагових векторів.

2.3 Новий метод зв'язку сусідніх нейронів мережі

Як вже було сказано, решітка класичної моделі самоорганізується нейронної мережі Кохонена має плоску прямокутну або гексагональную топологію (рисунок 2.2).



а) б)

Рисунок 2.2 – Решітки класичної моделі СОКК: а) прямокутна; б) гексагональна

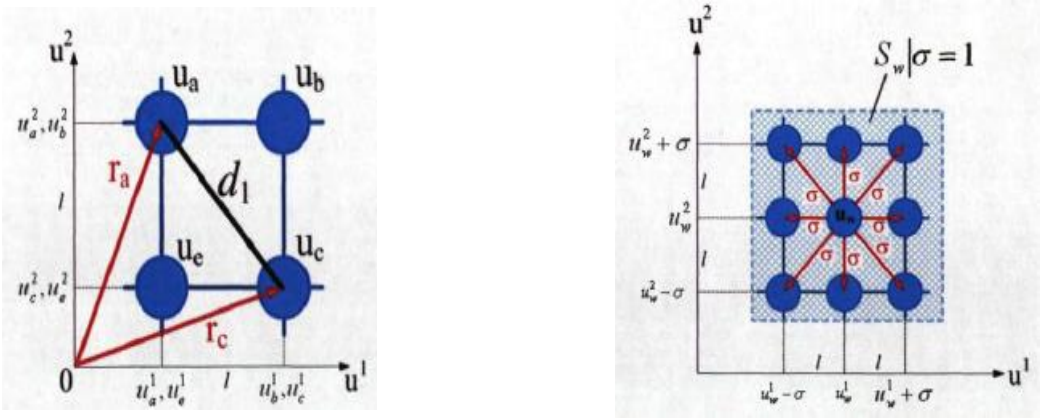
У даній роботі використовуються нейронні мережі з прямокутною топологією решітки мережі, тому далі будуть розглядатися тільки такі решітки. Необхідно нагадати, що нейронна мережа $U_N = \{u_1,...,u_N\}$,розташована в обчислювальної області $U \in R^2$, складається з нейронів $u_i \in U$, які мають свої координати $u_i = (u_i^1, u_i^2)$в системі координат (u^1, u^2) обчислювальної області U, де R – множина дійсних чисел, $i = 1,..., N$, N – кількість нейронів мережі.

У класичній моделі СОКК решітка мережі UN рівномірна, тобто відстань між нейронами в просторі (u^1, u^2) по вертикалі і по горизонталі однакове і дорівнює кроку l мережі. Цей крок приймається рівним 1 (l = 1) і використовується в даній роботі. Для визначення відстані між нейронами в такий решітці допускається будь-яка метрика. Так, відстань d між двома нейронами в решітки мережі класичної моделі розраховується як евклідова відстань між радіус-векторами ra і rc, проведеними з початку координат в решітці до цих нейронів:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.6}$$

а в область сусідства Sw заданого радіуса a для нейрона-переможця uw потрапляє безліч нейронів, що знаходяться від нього на відстані σ:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.7}$$



а) б)

Рисунок 2.3 – Відстань між нейронами та область сусідства в класичній СОКК: а) відстань між нейронами; б) область сусідства радіусу 1

У даній роботі пропонується наступний метод установки латеральних

зв'язків між нейронами в мережі і визначення розмірів топологічних областей сусідства.

Для кожного нейрона в решітці мережі створюється список його найближчих сусідів. Такими найближчими сусідами для нейрона ui, в радіусі $\sigma = 1$ є тільки 4 нейрона (рисунок 2.4), два з яких (ua і uc) знаходяться в тому ж стовпці, але в сусідніх по вертикалі і горизонталі рядках, а інші два нейрона (ub і ue) знаходяться в тому ж рядку, але в сусідніх зліва і справа шпальтах від розглянутого нейрона ui,. Таким чином, область сусідства радіуса $\sigma = 1$ для кожного нейрона ui в мережі з кроком l = 1 позначимо як:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.8}$$

В радіус сусідства $\sigma = 0$ потрапляє тільки сам нейрон ui, що відповідає значенням l=0.

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.9}$$

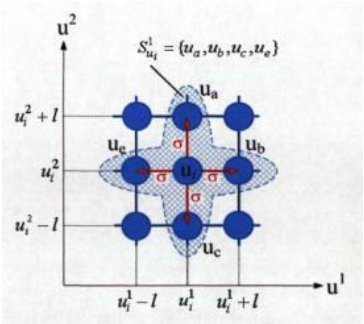


Рисунок 2.4 – Сусіди нейрона в радіусі $\sigma = 1$

Для пошуку сусідніх нейронів, що знаходяться в решітці мережі в радіусі сусідства σ, необхідно виконати σ кроків, на кожному з яких знайти найближчих сусідів по вертикалі і горизонталі (28) для вже знайдених

нейронів-сусідів на попередньому (σ-1) кроці:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.10}$$

В (30) для розрахунку S_{ui}^{σ} використовуються нейрони-сусіди і з області сусідства, яка також розраховується за формулою (30). Рекурсивний пошук нейронів-сусідів припиняється, при σ<1. При цьому, для σ = 1 безлічі нейронів-сусідів визначаються за формулами (28) і (29), відповідно.

Тоді для σ = 1, згідно (28) (рисунок 2.4):

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.11}$$

де σ=2:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.12}$$

де σ=3:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.13}$$

Тоді, враховуючи (31-33) також справедливо:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.14}$$

Список найближчих сусідів для кожного нейрона (28) задається на

стадії ініціалізації нейронної мережі і в процесі навчання не змінюється.

Запропонований метод установки латеральних зв'язків між нейронами і визначення розмірів топологічної області сусідства будь-якого радіуса не змінює структуру алгоритму самоорганізації Кохонена (7), а вбудовується в нього за рахунок зміни ваг латеральних зв'язків (12) між нейронами в решітці мережі.

Дана стратегія використовується в моделі самоорганізується нейронної мережі із замкнутою гратами, розробленої в атестаційній роботі.

2.4 Нові моделі СОКК із замкнутими гратами для усунення граничного ефекту

У даній роботі пропонується до використання варіант структури решітки нейронної мережі, яка була названа замкнутою. шляхом з'єднання країв решітки мережі з прямокутною топологією за наступним алгоритму пропонується вирішувати проблему "граничного ефекту" для СОКК.

На першому кроці алгоритму виробляється трансформація плоскою решітки в циліндр. При цьому з'єднуються один з одним її лівий і правий краю. Крайні нейрони в однойменних рядках решітки мережі з'єднуються між собою новими латеральними зв'язками, так що відстань між ними в решітці стає рівними кроці l.

Нейрони $u_i \in U_N$ плоскою решітки мережі, що належать її лівому краю, матимуть таку координату по осі u1 в системі координат обчислювальної області U:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.16)$$

а нейрони правого краю такої решітки

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.16}$$

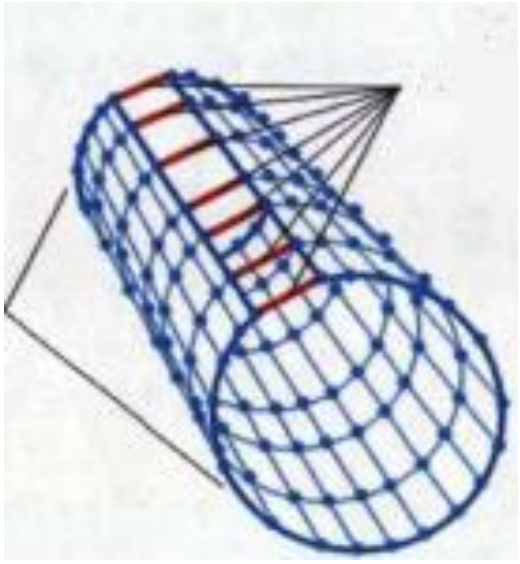


Рисунок 2.5 – З'єднання країв решітки мережі для створення циліндричної решітки

За рахунок з'єднання граничні нейрони на лівому і правому краях решітки мережі перестають існувати і, набуваючи сусідів, стають повноправними учасниками в алгоритмі навчання Кохонена і сусідами один для одного в радіусі $\sigma = 1$.

Нейрони $u_i \in U_N$, що належать верхній підставі такої циліндричної решітки, будуть мати наступну координату по осі u2 в системі координат обчислювальної області U:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.17}$$

а нейрони нижньої основи циліндричної решітки:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.17}$$

На другому кроці алгоритму необхідно з'єднати нейрони на утворилися підставах циліндра. Для цього пропонується дві схеми з'єднання:

Назва схеми: "Кожен з кожним" (рисунок 2.6). Всі нейрони кожного підстави циліндра решітки з'єднуються між собою єдиної латеральної зв'язком так, що стають сусідами один для одного в радіусі сусідства $\sigma = 1$ в межах кожного полюса.

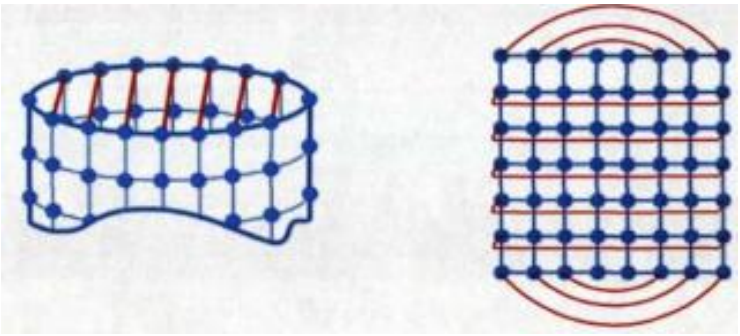


Рисунок 2.6 – З'єднання нейронів основахствореного циліндра замкнутої решітки мережі по схемі "Кожен з кожним"

Області сусідства для кожного нейрона в такий решітці мережі визначаються. Область сусідства радіуса $\sigma = 1$ для нейронів мережі з такою гратами буде складатися з об'єднання кількох множин

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{2.18}$$

де $S_{ui}^{\backprime 1}$. - безліч всіх нейронів верхнього підстави циліндрично решітки мережі:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.19)$$

$S_{ui}^{``1}$. - безліч всіх нейронів нижньої основи циліндричної решітки мережі:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.20)$$

Області сусідства для кожного нейрона в такий решітці мережі визначаються. Область сусідства радіуса σ = 1 для нейроні мережі з такою гратами буде складатися з об'єднання кількох множин:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.22)$$

де $S_{ui}^{`1}$ - нейрон-сусід на верхній основі циліндричної решітки мережі:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.23)$$

де $S_{ui}^{``1}$ - нейрон-сусід на нижній основі циліндричної решітки мережі:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \qquad (2.23)$$

На рисунку 2.8 наведені приклади областей сусідства Su1 і Su2 двох різних радіусів сусідства для двох нейронів-переможців 1 та 2 відповідно, мають однойменні координати в гратах мереж класичної (зліва) і розробленої моделей.

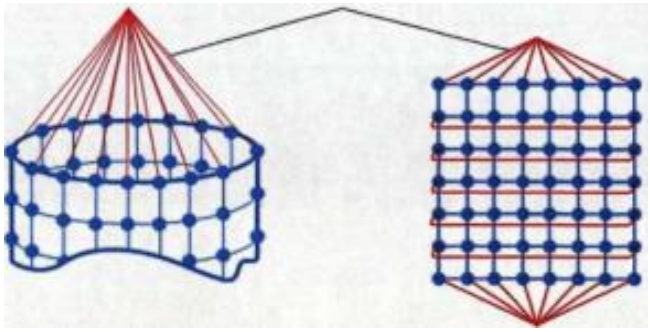


Рисунок 2.8 – Області сусідства різних радіусів на решітках різних моделей мереж

На ілюстрації (рисунку 2.9) побудовані графіки величини асиметрії латеральних зв'язків δ_i для розглянутих мереж при одній і тій же зафіксованії ітерації навчання і радіусі навчання $\sigma = 1$

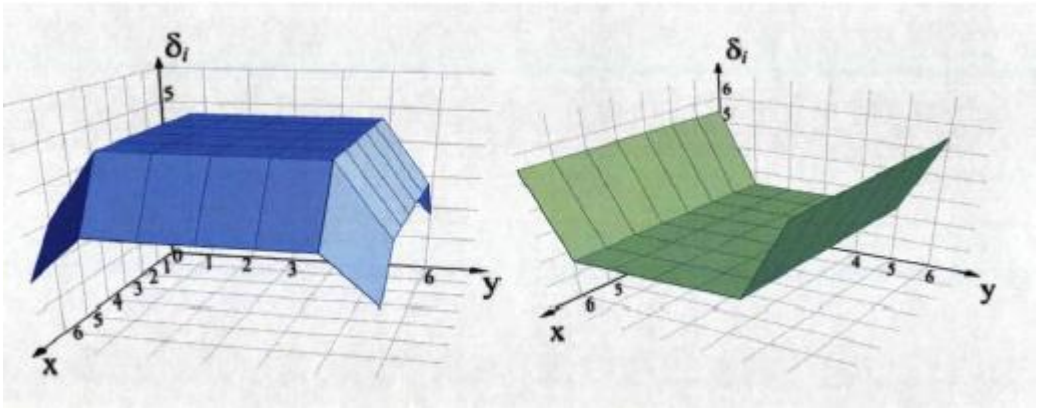


Рисунок 2.9 – Характеристика симетричності латеральних зв'язків для СОКК

Як видно в класичній моделі СОКК для нейронів на всіх кордонах мережі не вистачає сусідів, тому характеристика δ_i зменшується по периметру решітки мережі. Нейрони ж, що знаходяться на відстані від кордону решітки на відстані σ>1, мають рівні значення характеристики δ_i.

Замкнута решітка нейронної мережі збільшує кількість сусідів граничної мережі нейронів. Для схеми такої мережі "Кожен з кожним" нейрони на полюсах решітки навіть перенасичені сусідами. на інших 2-х

краях мережі граничний ефект вирівняний.

Схема "Протилежні" є практично ідеальною, з точки зору характеристики δ_i, і за допомогою неї здійснюється майже повне балансування граничного ефекту.

3 МОДЕЛЮВАННЯ МОДИФІКОВАНОЇ МЕРЕЖІ КОХОНЕНА

3.1 Оцінка точності і якості навчання мережі із замкнутими решітками

Оцінка точності і якості апроксимації вхідного простору аналізованих даних для розробленої моделі замкненої мережі виконувалося на основі двох критеріїв:

а) критерію точності навчання мережі - помилки квантування;

б) критерію якості навчання - ентропії мережі. Для досліджень використовувалися мережі класичної та розробленої моделей, навчених за інших рівних умов:

- однакові розміри мереж: v = 37 нейронів, h = 21 нейрон, N = 999 нейронів;
- однакові початкова швидкість навчання a0 = 0,8 і початковий радіус навчання σ_t = 10 мереж. Однакові для мереж функції зміни швидкості (9) і радіусу сусідства (11);
- функція сусідства для обох мереж гауссовского (12) типу;
- вектори вхідного набору даних перед навчанням мережі нормалізуються (19).

В якості моделі мережі Кохонена використана мережу з некомерційного безкоштовного пакета програм "SOM PAK", створеного в Лабораторії обчислювальної техніки та інформатики Хельсінского технологічного університету.

Навчання мереж проводилося на одному і тому ж вхідному наборі даних, що представляє собою 2000 випадкових точок тривимірного простору XYZ, координати яких розподілені щодо початку координат (0, 0, 0) по нормальному закону з нульовим математично очікуванням і среднеквадратическим відхиленням, рівним 1.

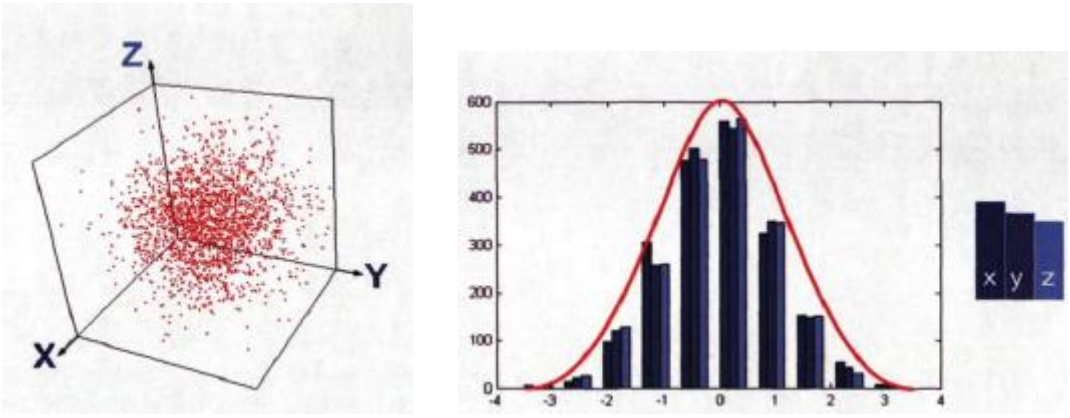


Рисунок 3.1 – Вхідний набір даних для мереж

Помилка квантування мережі як оцінка точності навчання. Метою навчання мережі з самоорганізацією на основі конкуренції нейронів вважається таке впорядкування нейронів (підбір значень їх ваг), яке мінімізує значення очікуваного спотворення, що оцінюється похибкою апроксимації вхідного вектора x = (x1, x2, ..., $x_м$) (м - кількість компонент вхідного вектора), значеннями ваг нейрона-переможця в конкурентній боротьбі. При p вхідних векторах x і застосуванні евклідової метрики ця похибка, звана також помилкою 1 P квантування.

Порівняння результатів роботи мереж класичної та розробленої моделей за критерієм помилки квантування (14) проводилося при різному кількості епох навчання (10, 30, 50, 100 і 200 епох).

Результати проведених до сліджень наведені в таблиці 3.1 і показані на (рисунку 3.2). Обидві схеми замкнутої мережі ("протилежні" і "кожен з кожним") представлені одним графіком моделі в цілому, тому що за критерієм помилки квантування ці дві схеми дають однакові результати.

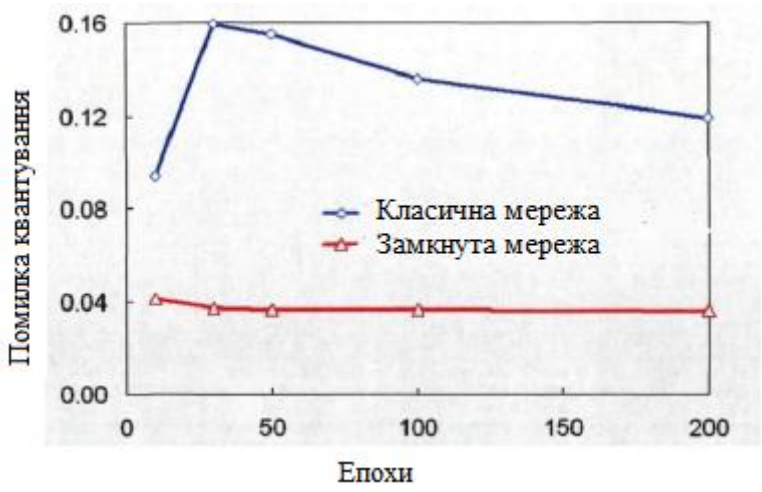


Рисунок 3.2 – Графік помилки квантування мереж

Таблиця 3.1 – Результати дослідження помилки квантування мереж

Кількість епох	Похибка квантування		
	Класична мережа E_{q1}	Замкнута мережа E_{q2}	Відношення $\frac{E_{q1}}{E_{q2}}$
10	0.094	0.041	2.293
30	0.159	0.037	4.251
50	0.155	0.036	4.270
100	0.136	0.036	3.747
200	0.119	0.036	3.324
Середнє значення:			3.577

На початкових етапах навчання вплив кожного вхідного вектора на нейронну мережу настільки велике (за рахунок відносно великих початкових значеннях швидкості a0 і радіусу сусідства навчання мережі), що під значення його компонент прагнуть підлаштуватися майже всі нейрони (Їх вагові вектори) мережі. Це призводить до зростання помилки квантування у мережі Кохонена класичної моделі. З кожним новим вхідним вектором мережа прагне підлаштуватися під значення його компонент, практично повністю "забуваючи" про існування попереднього вхідного вектора.

Однак з кожною епохою навчання вплив вхідних векторів на нейронну мережу слабшає (зменшуються швидкість навчання і радіус сусідства). Все меншій кількості нейронів в мережі "дозволяється" підлаштовуватися під значення компонент вхідного вектора. На останніх етапах навчання відбувається тонке налаштування нейронної мережі під вхідні дані, коли свої ваги адаптує під вхідний вектор тільки нейрон-переможець.

При малій кількості епох навчання (графіки помилки квантування мереж при кількості епох навчання менше 50) нейронних мереж не вистачає часу для навчання. Швидкість навчання і радіус сусідства швидко зменшуються, і мережі грубо аппроксимирует вхідні дані, від чого помилка квантування має великі значення. При проведенні більшого кількості епох навчання, швидкість навчання і радіус сусідства приймають невеликі значення, що дозволять виробляти більш тонку підстроювання ваг нейронів під вхідні вектори. При цьому графіки помилки квантування мереж стають більш пологими.

У таблиці 3.1 наведені розрахунки відхилень в значеннях помилки квантування на різних епохах. Середнє значення відносини – помилки квантування для двох мереж дорівнює приблизно 3.6. Цей критерій показує, що самоорганізується мережу розробленої моделі в 3.6 рази точніше аппроксимирует простір вхідних даних, ніж мережу класичної моделі цього типу.

3.2 Ентропія мережі як оцінка якості навчання

У своїй класичній праці Клод Шеннон виклав основи теорії інформації. У цій роботі було введено поняття "ентропії" за аналогією з ентропією в термодинаміки. Ентропія в теорії інформації представляють собою міру середнього обсягу інформації, яке може містити в собі повідомлення:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{3.1}$$

Також справедливі наступні твердження:

- H(Y) = 0 тоді і тільки тоді, коли pi = 1 для деякого i -го значення змінної, а для всіх інших значень ймовірність дорівнює нулю. Це нижня межа ентропії відповідає відсутності невизначеності;

- H(Y) = log2C з тоді і тільки тоді, коли pt =1/c для всіх i (тобто якщо ймовірність всіх значень змінної Y різновірогідні). Таким чином, ентропія H(Y) може набувати таких значень:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{3.2}$$

Для самоорганізованої нейронної мережі ентропія – це ступінь рівномірності розподілу помилки квантування по нейронам мережі. Чим рівномірніше помилка квантування розподілена по кожному нейрону мережі, тим вище значення цього критерію.

Максимальна кількість кластерів, які мережа може виявити у вхідних даних дорівнює кількості її нейронів – N.

Кожен нейрон мережі може виступати центром будь-якого кластера для деякої кількості l вхідних векторів x_j вхідного набору аналізованих мережею даних. Тобто вектор ваги w_n такого нейрона ип буде самим близьким, в сенсі евклідової метрики, до цих вхідних векторів. Тоді точність апроксимації кожним нейроном потрапили в нього за все час навчання T вхідних векторів позначимо як:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{3.3}$$

У кожен нейрон мережі за весь час навчання може потрапити будь-яка кількість векторів. Однак для двох нейронів, з однаковою кількістю потрапивших в них вхідних векторів, значення Еп може виявитися різним.

Для проведення порівняння наскільки один нейрон мережі аппроксимує вхідні вектори точніше іншого нейрона, необхідно нормалізувати їх точності по загальній точності апроксимації всього вхідного набору аналізованих даних всієї нейронною мережею:

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{3.4}$$

де К – кількість векторів всього вхідного простору аналізованих даних.

$$L = \frac{3,6Q}{c(t_1 - t_2)}, \tag{3.5}$$

де N – кількість нейронів мережі.

Порівняння результатів роботи мереж класичної та розробленої моделей за критерієм ентропії проводилося при різній кількості епох навчання (10, 30, 50, 100 і 200 епох). Результати проведених досліджень наведені в Таблиці 3.3 і на (рисунок 3.3). Обидві схеми замкнутої мережі ("Протилежні" і "кожен з кожним") представлені одним графіком моделі в цілому, тому що за критерієм ентропії ці дві схеми дають однакові результати.

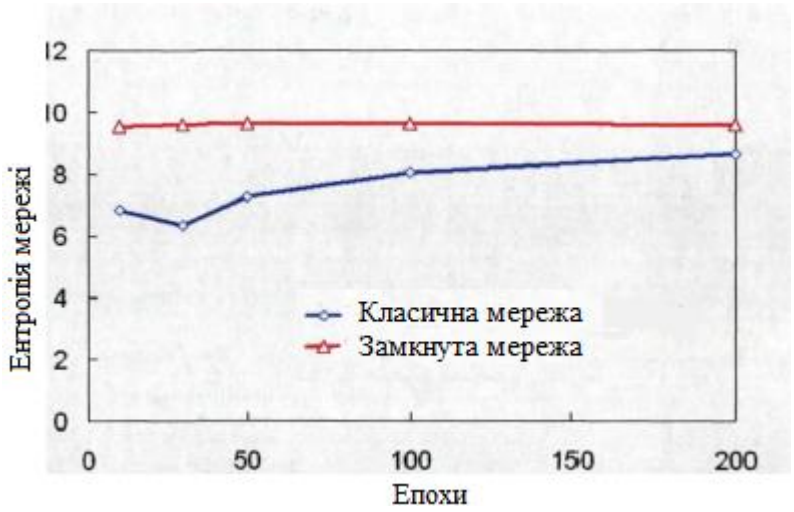


Рисунок 3.3 – Графік помилки квантування мереж

Таблиця 3.2 – Результати дослідження ентропії мереж

Кількість епох	Похибка квантування		
	Класична мережа H_1	Замкнута мережа H_1	Відношення, % $\left(\frac{H_2}{H_1}-1\right)\cdot 100\%$
10	6.811	9.516	40
30	6.311	9.588	52
50	7.247	9.617	33
100	8.049	9.632	20
200	8.632	9.606	11
Середнє значення:			31

Середнє значення показника $\left(\frac{H_2}{H_1}-1\right)\cdot 100\%$ в таблиці 2 показує, що самоорганізована мережа замкнутої моделі на 31% якісніше аппроксимує вхідний простір аналізованих даних в порівнянні з класичною моделлю мережі цього типу. Збільшення точності і якості апроксимації розробленої мережею вхідного простору даних відбувається через збільшення кількості латеральних зв'язків між нейронами і більш рівномірного розподілу по

мережі векторів вхідного простору. На рисунку 2.13 наведені карти частот для мереж класичної та розробленої замкнутої моделей.

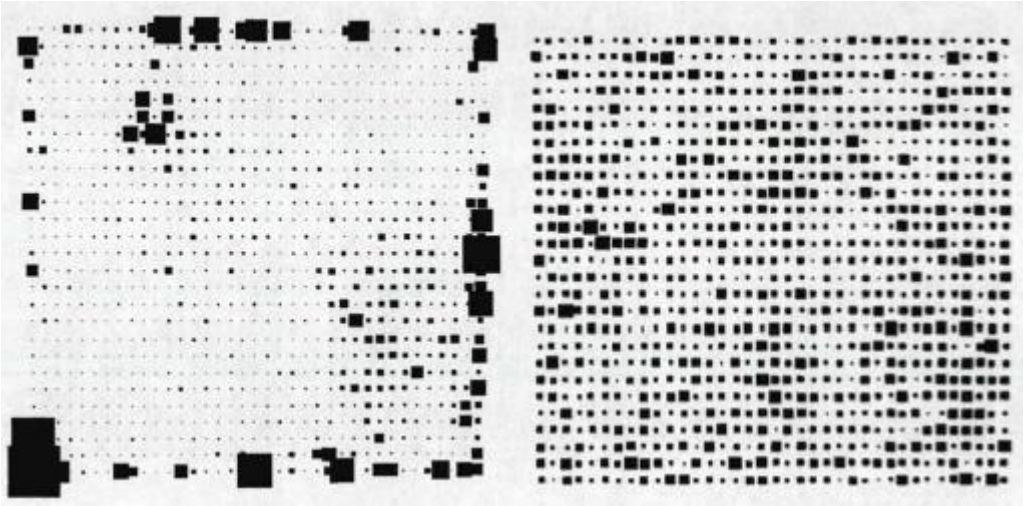


Рисунок 3.4 – Карти частот для класичної і замкнутої моделей мереж

Розмір кожної ділянки на картах пропорційний кількості вхідних векторів, що потрапили в нейрон мережі. Кожна ділянка на карті відповідає певного нейрона в мережі. На карті класичної моделі видно, що вхідні вектори в процесі навчання захоплювалися тільки частиною нейронів мережі (ділянки великих розмірів). Решта нейрони (ділянки малих розмірів) практично не вигравали для вхідних векторів. У замкнутої мережі майже всі нейрони мережі брали участь в адаптації своїх вхідних векторів, про що свідчить майже однаковий розмір ділянок на карті частот замкнутої мережі.

Для підтвердження цього твердження на рисунку 2.14 побудовано 2 гістограми щільності розподілу нормализованной помилки квантування по нейронам мережі з різною точністю.

Вище було зазначено, що розміри нейронних мереж (v•h = N), які беруть участь в дослідженні однакові і рівні N = 999 нейронів. справа на гистограммі показано, що приблизно 500 (близько половини) нейронів класичної моделі і тільки 80 нейронів моделі замкненої мережі мають найвище значення помилки квантування з розрахованих.

Таке значення помилки квантування присвоєно нейронам мереж, які за весь час навчання жодного разу не ставали переможцями ні для одного вхідного вектора.

Таким чином, в розробленій моделі самоорганізується нейронної мережі відбувається більш рівномірний розподіл вхідних векторів по нейронам мережі в порівнянні з класичною моделлю мережі Кохонена.

В Таблиці 3.3 наведені результати порівняння точності і якості аналізу даних цими мережами по відношенню один до одного.

Таблиця 3.3 – Результати дослідження помилки квантування і ентропії для замкнутих мереж

Кількість епох	Похибка квантування моделей			Ентропія моделей		
	E_{Δ}	$E_{\square}$	$\frac{E_{\Delta}}{E_{\square}}$	H_{Δ}	$H_{\square}$	$\left(\frac{H_{\square}}{H_{\Delta}}-1\right)\cdot 100\%$
10	0.041	0.041	1	9.32	9.52	2%
30	0.034	0.0374	0.91	9.49	9.59	1%
50	0.033	0.0363	0.91	9.50	9.62	1%
100	0.0315	0.0363	0.877	9.51	9.63	1%
200	0.0309	0.0358	0.863	9.53	9.61	0.8%
Середнє:			0.91	Середнє:		1%

За значеннями відхилень помилки квантування і ентропії для мереж можна зробити висновок, що за характеристиками точності і якості обидві мережі майже не поступаються одна одній. Перевага моделі мережі з прямокутної топологією решітки полягає в більш простій реалізації в порівнянні з трикутною топологією, а також в можливості побудови наочних плоских карт-розгорток замкнутої мережі.

Відомі проблеми СОКК класичної моделі, такі як "граничний ефект "і" мертві нейрони "впливають на точність і якість аналізу даних. Для подолання

цих проблем зі збереженням важливих властивостей класичної моделі СОКК запропонована нова модель мережі цього типу, в решітці якої нейрони на кордонах з'єднуються між собою. Нова модель в кілька разів точніше і якісніше аналізує вхідний набір даних, в порівнянні з класичною моделлю СОКК.

3.3 Вибір інструментальних засобів

Комплекс програм, реалізований для виконання операцій аналізу даних із застосуванням самоорганізуються нейронних мереж Кохонена, складається з наступних компонентів:

- динамічно приєднуються бібліотека з відкритим інтерфейсом, що реалізує алгоритм навчання самоорганізуються нейронних мереж як класичної моделі карти Кохонена, так і її модифікованих варантів;

- додаток, розроблений в універсальному табличному редакторі Microsoft Excel, що використовує розроблену бібліотеку та забезпечує інтерфейс користувача для підготовки даних до аналізу, а також для візуального відображення та навігації по картах навчених нейронних мереж;

- динамічно приєднана бібліотека з відкритим інтерфейсом, реалізована в середовищі "C ++ Builder" і призначена для вбудування в системи підтримки прийняття рішень і реалізує роботу самоорганізуються нейронних мереж Кохонена;

- бібліотека підтримує можливість встановлення користувачем форм мережі (плоскій або замкнута), її розмірів, початкових параметрів ініціалізації та кількості навчання. Для задання вхідних даних для бібліотечних методів і в якості вихідних даних використовуються файли текстового формату, як формату, що підтримується в більшості додатків.Алгоритми розробленої бібліотеки були застосовані в додатку, розробленій в універсальному табличному редакторі Microsoft Excel (MS Excel). Приложение представляє собою електронну книгу, розділену на

листки:

- лист входных даних;
- лист результатів.

Для визову методів динамічної бібліотеки та візуального відображення карт обучених СОКК використовується технологія макросів, яка в Microsoft Excel підтримується за допомогою програмного інтерпретатора Visual Basic for Applications.Лист вхідних даних електронної книги містить додатки, в рядках яких перераховані вектори вхідного набору даних, а в колонках - компоненти цих векторов. Лист вхідних даних призначений для здійснення первинної підготовки вхідних даних для аналізу. Це дозволяє очистити вхідні вектори від пропусків у значеннях компонент, або заповнити пропуски службовими символами "x". Виконати фільтрацію вхідних векторів, їх сортування, відбір необхідних колонок таблиць, як кінцевих компонентів векторов.Моделювання роботи мережі здійснювалося наступним чином: подана вхідна вибірка даних для 7 різних компонент, у кожного з цих компонентів є n-е кількість ознак; далее відбувається навчання мережі; таким чином було проведено моделювання не тільки класичного варіанта навчання карт Кохонена, але і її модифікованих варіантів.

Тестовий набір даних представлений в таблиці 3.4.

Таблиця 3.4 – Тестовий набір даних

№	Назва компоненти (параметра)	Опис параметра
1	Параметр 1	4-12
2	Параметр 2	12-50
3	Параметр 3	1,2
4	Параметр 4	1-5
5	Параметр 5	1-16
6	Параметр 6	0-100
7	Параметр 7	1-12

Результати аналізу даних за допомогою нейронної мережі:Параметри обучаемой мережі:

- тип мережі: класична мережа Кохонена, ікосаэдр, модель з замкнутою решіткою ";

- розміри мережі: v = 50 нейронів, h = 30 нейронів ;.N = v * h = 1500 нейронів;

- початкова швидкість навчання a = 0,75 зменшується; процес навчання за лінейним законом;

- початковий радіус; навчання r = 25 зменшується в процесі навчання за лінійним законом;

- функція сусідства гауссовского типу.Проведено100 епох навчання; ошибка квантования: складала: E = 0,011. Результати навчання мережі у вигляді розгортання показані на рисунку 3.5.

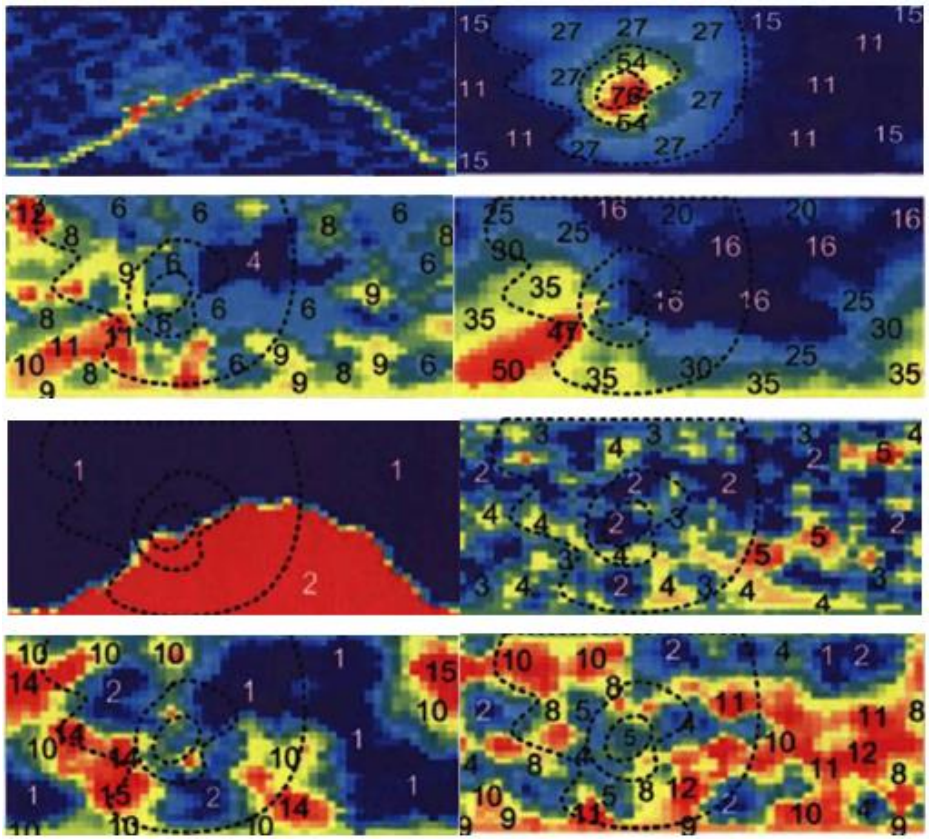


Рисунок 3.5 – Результати моделювання для моделі з замкнутою решіткою

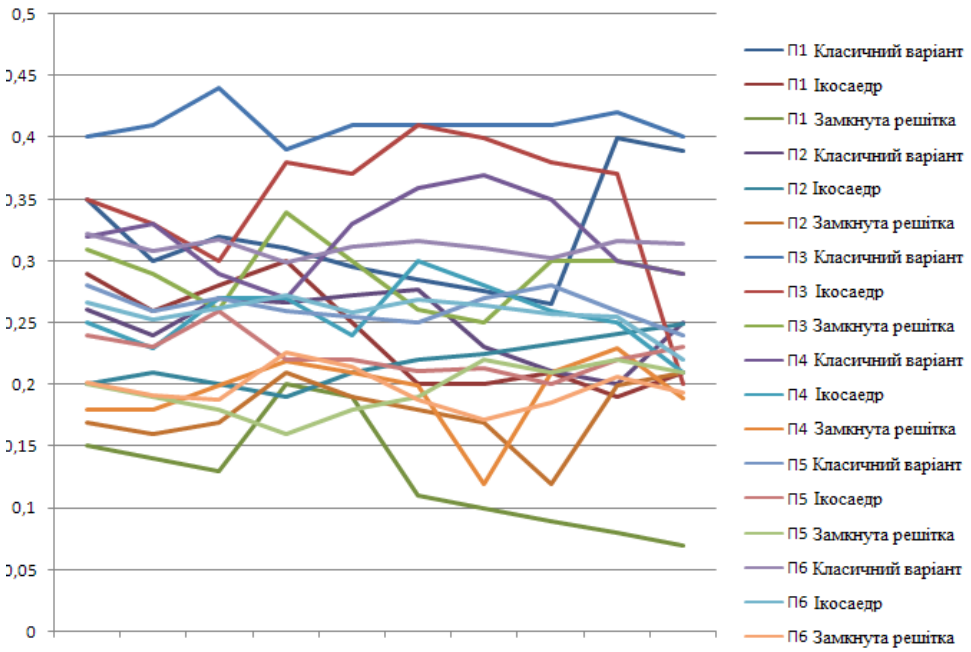


Рисунок 3.6 – График залежності величини помилки від вибраного варіанта навчання для всіх параметрів

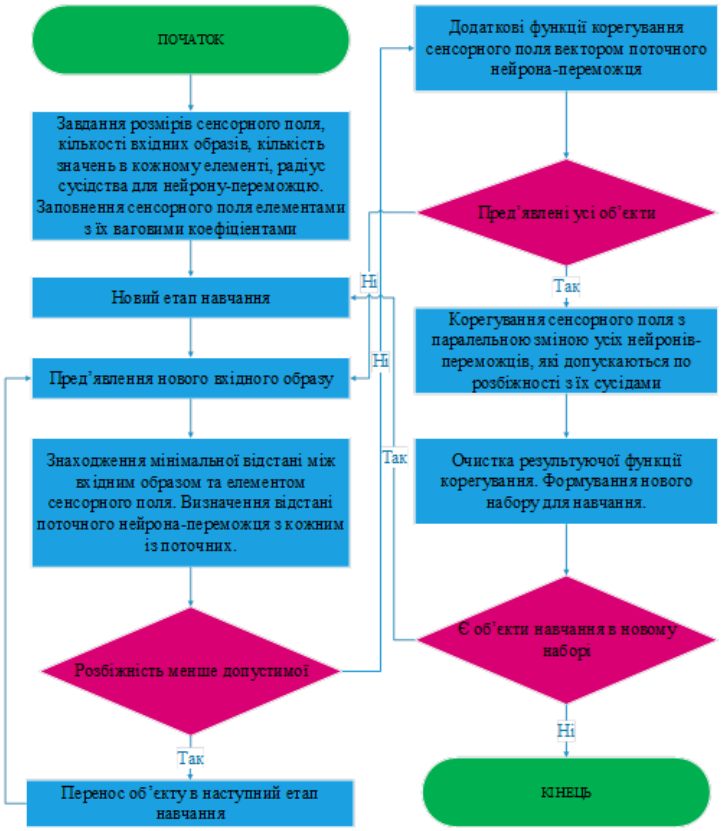


Рисунок 3.7 – Метод навчання модифікованої карти Кохонена

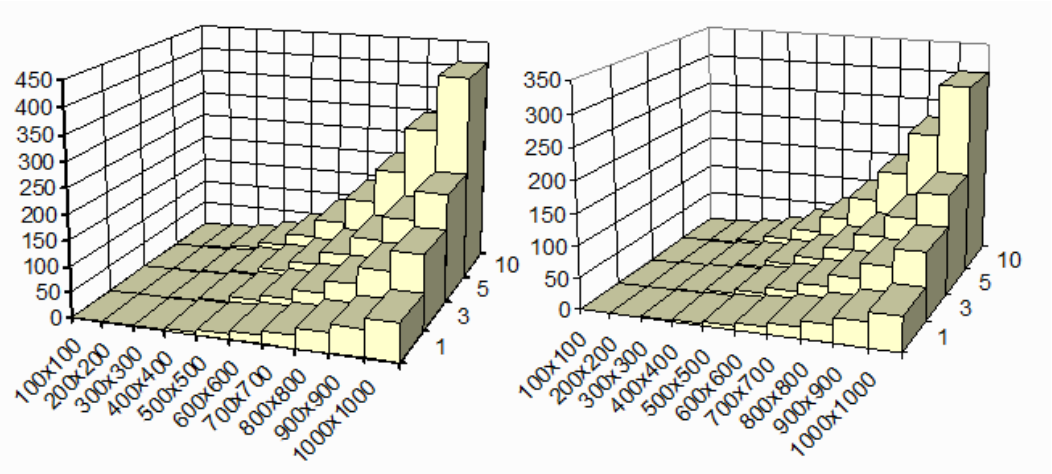


Рисунок 3.8 – Порівняльний аналіз навчання традиційного та модифікованого варіантів навчання карт Кохонена

Карта кластерів розділена на 2 половини красно-жовтого: границя. Це виникло за рахунку наявності вхідного набору векторів класифікуючого параметра1, який приймає тільки два значення.

На рисунку 3.6 представлені результати експериментів для всіх параметрів та 3х алгоритмів навчання мереж Кохонена:

- класичний варіант
- ікосаэдра
- модели з замкнутою решеткою.

Отримані результати дозволяють сказати, що найкраща ефективність показала модель Кохонена з замкнутою решіткою.

ВИСНОВКИ

Класична модель нейронної мережі, що самоорганізується, запропонована Кохонном, надає можливості кластерного аналізу багатомірних даних і є засобом візуального відображення характерів їх розподілу на двомірному топологічному картці. Значні проблеми СОКК класичної моделі, такі як "граничний ефект" і "мертві нейрони" впливають на точність і якість аналізу даних. Для подолання цих проблем із збереженням важливих властивостей класичної моделі СОКК розроблені нові модифіковані варіанти мережі цього типу, в решітці яких нейрони на кордонах з'єднуються між собою.

При виконанні даної роботи були вивчені карти Кохонена, що самоорганізуються. Розглянуті як класичні, так і модлифіковані варіанти навчання мережі. Розроблене програмне забезпечення, реалізує дані алгоритми. Проведені дослідження ефективності навчання модифікованих мереж Кохонена. Найкращі результати показали метод навчання мережі Кохонена з замкнутою решіткою.

Встановлено зв'язки між способами з'єднання нейронів на кордонах решітки мережі Кохонена та ефективністю апроксимації нею вхідного набору аналізованих даних. Розроблено метод установки зв'язку між нейронами в решітці замкнутої мережі і описана його математична модель, яка спрощує алгоритм встановлення розмірів топологічних областей сусідів.

Розроблена модель нейронної мережі вирішує проблеми граничного ефекту і появи "мертвих нейронів", які властиві класичній моделі мережі цього типу, а також збільшує точність і якість апроксимації аналізованих даних.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Розенблатт Ф. Принципы нейродинамики. Персептрон и теория механизмов мозга: Пер. с англ. – М.: Мир,1965.-332 с.

2. Demuth H., Beale M. Neural Network Toolbox. For Use with MATLAB. The MathWorks Inc., 1992-2000.-415 с.

3. Нейронные сети. STATISTICA Neural Networks. М. Горячая линия – Телеком, 2000.-360 с.

4. Principe J.C., Euliano N.R., Lefebvre W.C. Neural and Adaptive Systems. Fundamentals Through Simulations.- New York: John Wiley Sons Inc., 2000.-403 с.

5. Luo F-L., Unbehauen R. Applied Neural Networks for Signal Processing.- Cam-bridge University Press, 1998.-380 с.

6. О.Г.Руденко, Е.В. Бодянский. Искусственные нейронные сети. – Харьков, СМИТ, 2005.-408 с.

7. Галушкин А.И. Теория нейронных сетей. Кн.1: Учебное пособие дл вузов – М.: ИПРЖР, 2000. – 416 с.