

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ КЛАСТЕРІЗАЦІЇ БАГАТОВИМІРНИХ ДАНИХ МЕТОДОМ КАРТ КОХОНЕНА

При реалізації алгоритму кластеризації даних з використанням нейронних мереж, що самоорганізуються, було продемонстровано двовимірне відображення багатовимірних даних, використовуючи клас функцій plotsom. Були розглянуті способи розрахунку відстані від об'єкта до центру кластера, та вплив обраних функцій на кінцеве розміщення кластерів. Проаналізовані особливості задавання та роботи карт Кохонена.

Розвиток обчислювальної техніки призводить до появи нових інформаційних технологій, зокрема останнім часом стрімко розвиваються методи обробки інформації, засновані на нечіткій логіці і нейронних мережах.

Кластерний аналіз інформації є одним з важливих етапів, що виконуються при розв'язанні практично будь-якої задачі, пов'язаної з аналізом багатовимірної інформації.

Серед багатьох різновидів нейронних мереж для даного типу задач ефективністю виділяється клас мереж, які називаються картами Кохонена, що самоорганізуються. Їх особливістю є адаптивність, двовимірне відображення багатовимірних даних, топологічна близькість одержуваних кластерів і наочність вихідної інформації. Карты Кохонена [1] використовуються, в першу чергу, для візуалізації та початкового («розвідувального») аналізу даних. Нейронні мережі даного типу часто застосовуються для вирішення самих різних завдань, від відновлення пропусків у даних до аналізу даних і пошуку закономірностей. У даній роботі розглянуто принципи функціонування та деякі аспекти використання карт, що самоорганізуються [2].

Алгоритм функціонування карт, що самоорганізуються (Self Organizing Maps – SOM) являє собою один з варіантів кластеризації багатовимірних векторів. Важливою відмінністю алгоритму SOM є те, що в ньому всі нейрони (вузли, центри класів) впорядковані в деяку структуру (зазвичай

двовимірну сітку). При цьому в ході навчання модифікується не тільки нейрон-переможець, але і його сусіди, але меншою мірою. За рахунок цього SOM можна вважати одним з методів проектування багатовимірного простору в простір з більш низькою розмірністю. При використанні цього алгоритму вектора, схожі у вихідному просторі, виявляються поруч і на отриманій карті [3].

Алгоритм складається з наступних кроків: спочатку задаються параметри нейронної мережі (кількість нейронів по горизонталі, кількість нейронів по вертикалі), вибір топології, функція розрахунку відстані між центром кластера і точкою даних, кількість кроків для навчання, кількість сусідніх нейронів, які будуть зсуватися разом із змінною координат нейрона. Для зручності подальших розрахунків обчислюється кількість кластерів.

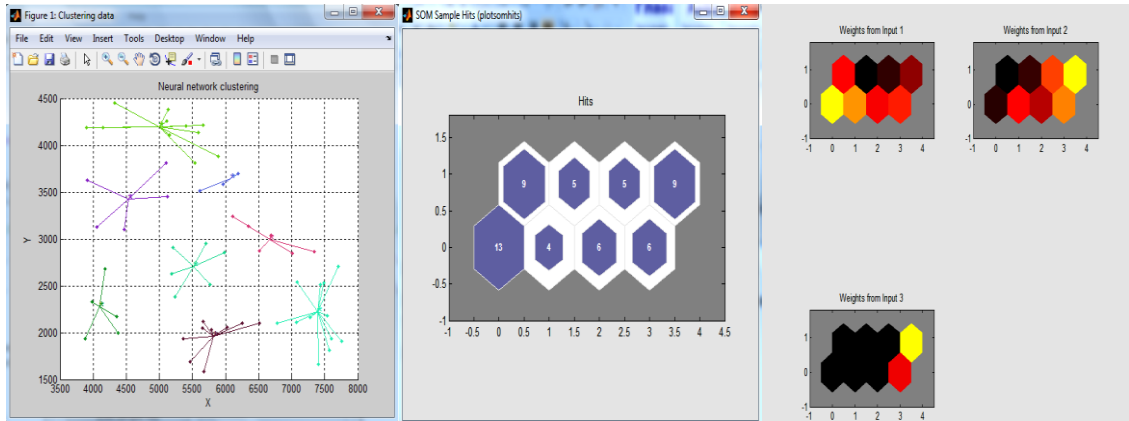
Після визначення всіх основних параметрів відбувається початкова ініціалізація мережі та її тренування на вхідних даних.

```
net = newsom(data, [dimension1
dimension2], Topology, Distance_func, ...
Steps, Initial_neighbor);
net = init(net);
net = train(net, data);
distances = dist(data', net.IW{1}');
[d, idx] = min(distances, [], 2);
```

За допомогою функції `newsom()` створюється карта Кохонена з заданими вхідними параметрами. Дана карта ініціалізується функцією `init(net)`. Тренування відбувається за допомогою стандартної функції `train(net, data)`, вхідними параметрами якої є новостворена карта і навчальна вибірка, в даному випадку вхідні дані. Для визначення приналежності точки до даного кластеру розраховується відстань від точки до центру кожного кластера. У результаті точка належить кластеру, відстань до якого виявилась найменшою. Після визначення відстаней створюється матриця приналежності кожної точки до конкретного кластера. Після знаходження центрів кластерів і визначення належності точок до даних кластерів

виводиться результат у вигляді отриманих центрів і ліній, що з'єднують точки даних з відповідним їм кластером.

Результат кластеризації нормалізованих даних з трьома параметрами представлений на рис.1 (а,б,в).

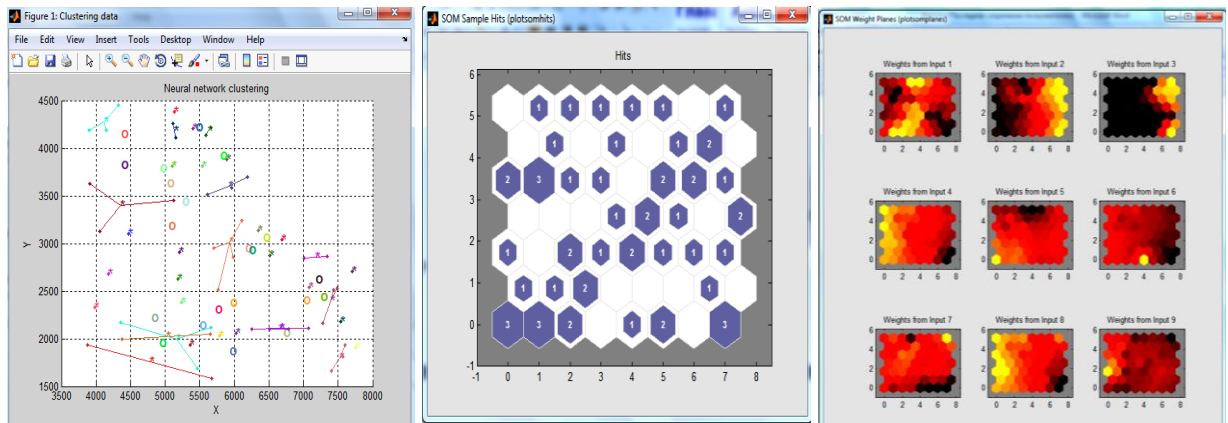


а – центри кластерів при кластеризації тримірних даних;
б – кількість влучень у нейрони при кластеризації тримірних даних;
в – вагові коефіцієнти нейронів при кластеризації тримірних даних.

Дослідивши отримані результати можна зробити висновок, що з метою отримання достовірних центрів кластерів при кластеризації багатовимірних даних методом карт Кохонена, слід використовувати якомога більше параметрів, що характеризують дані.

Однією з переваг використання карт, що самоорганізуються є можливість не здавати кількість кластерів до навчання. Дана нейронна мережа сама згрупує ознаки в необхідне число кластерів.

Результат кластеризації нормалізованих даних з дев'ятьма параметрами та розмірністю карти Кохонена $8 * 7$ представлений на рис.2 (а,б,в).



а

б

в

Рис. 2. а – центри кластерів при кластеризації;

б – кількість влучень у кожен нейрон; в – вагові коефіцієнти нейронів;

З даного експерименту видно, що 18 нейронів виявилися порожніми, отже розмірність карти можна зменшити мінімум на 18 нейронів і це не позначиться на якості кластеризації.

В роботі використано пакет Neural Toolbox, який містить функції для побудови, та навчання нейронних мереж, і кластеризації даних за допомогою карт Кохонена, розроблено програмне забезпечення для кластеризації багатомірних даних, перевірено працездатність програмного забезпечення на тестових даних.

Перелік літератури:

1. Кохонен Т. – Самоорганізуючі карти – М.:Бином, 2008. –659 с.
2. [http://ru.wikipedia.org/wiki Самоорганізующая карта Кохонена](http://ru.wikipedia.org/wiki/Самоорганізующая_карта_Кохонена)
3. [http://ru.wikipedia.org/wiki Нейронная сеть Кохонена](http://ru.wikipedia.org/wiki/Нейронная_сеть_Кохонена)