УДК 681.322

Ю.В. Стех, М.Е. Сардіх Файсал, М. В. Лобур, А.Б. Керницький
Національний університет "Львівська політехніка",
кафедра систем автоматизованого проектування

АЛГОРИТМ ПОШУКУ ОПТИМАЛЬНОЇ КІЛЬКОСТІ КЛАСТЕРІВ

Пропонується новий алгоритм пошуку оптимальної кількості кластерів для задач кластерного аналізу багатовимірних об'єктів. Центри кластерів знаходяться за допомогою нейроподібної процедури. Алгоритм не вимагає жодної початкової інформації про кількість і розміщення центрів кластерів. Наведені результати тестових досліджень.

Ключові слова – кластерний аналіз, багатовимірний об'єкт, алгоритм

The new algorithm of search of optimum amount of clusters is offered for the tasks of cluster analysis of multidimensional objects. Centers of clusters are search by neurolike procedure. An algorithm requires not a single initial information about an amount and placing of centers of clusters. The results of researches of tests are resulted.

Keywords – cluster analysis, multidimensional object, algorithm

Формулювання проблеми

Кластерний аналіз даних – це задача розбиття множини N об'єктів (вібірки об'єктів) на непересічні підмножини, які називаються кластерами, так, щоб кожен кластер складався з подібних об'єктів, а об'єкти різних кластерів істотно відрізнялися [1]. Кожен об'єкт представляється набором m своїх характеристик, які називаються ознаками. Ознаки можуть бути як числовими, так і нечисловими. Отже, кожен об'єкт представляється точкою в m-вимірному просторі xi $\in$ Rm, i= 1,2,…N. Кожен об'єкт характеризується відстанями до всіх решти об'єктів вибірки.

Типові завдання кластеризації такі:

- розуміння даних шляхом виявлення кластерної структури. Розбиття вибірки на групи подібних об'єктів дозволяє спростити подальше оброблення даних і прийняття рішень шляхом застосування до кожного кластера своїх методів аналізу;
- стиснення даних. Якщо початкова вибірка доволі велика, то можна зменшити її, залишивши по одному або по декілька найтиповіших представників від кожного кластера;
- виділення нетипових об'єктів, котрі не вдається приєднати до жодного кластера.

У першому випадку число кластерів намагаються зробити якомога меншим. У другому винадку важливо забезпечити високий ступінь подібності об'єктів в середині кожного кластера, а кластерів може бути довільна кількість. У третьому випадку найбільший інтерес становлять окремі об'єкти, які не вписуються в жоден кластер.

Формальна постановка задачі кластеризації

Нехай X – множина об'єктів

$$i = (xi1, xi2, \ldots xid) \in A \qquad (1)$$

в просторі ознак A, де i = 1,2,…N і кожен компонент xij $\in$ A є числовий або нечисловий атрибут.

Y – множина номерів кластерів. Такий формат вибірки відповідає матриці розмірністю N × d. Задана функція відстані ρ(xi, xj). Метою кластерного аналізу даних є розбиття вибірки на непересічні підмножини, які називаються кластерами так, щоб кожен кластер складався з об'єктів, які близькі по метриці ρ, а об'єкти різних кластерів істотно відрізнялися

$$X = C1 \cup C2 \cup \ldots \cup Ck \cup Coutliers,\ Ci \cap Cj = \varnothing \quad (2)$$

При цьому кожному об'єкту xi ∈ A приписується номер кластера yi ∈ Y.

Алгоритм кластеризації – це функція F : A → Y, котра довільному об'єкту xi ∈ A ставить у відповідність номер кластера yi ∈ Y. Множина Y в деяких випадках може бути відома заздалегідь. Здебільшого ставиться задача визначити оптимальне число кластерів, яке задовольняє певний критерій якості кластеризації. Кластеризація належить до задач навчання без учителя і істотно відрізняється від задач класифікації (навчання з учителем) тим, що кількість і початкове положення кластерів не задані.

Результат поділу заданої множини об'єктів на кластери залежить від таких факторів:

- не існує однозначно найкращого критерію якості кластеризації. Відомо багато евристичних критеріїв, а також алгоритмів, які не мають чітко вираженого критерію;
- число кластерів, як правило, невідоме заздалегідь і встановлюється відповідно до деяких суб'єктивних критеріїв;
- результат кластеризації істотно залежить від метрики ρ, вибір котрої, як правило, також суб'єктивний і визначається експертом.

Кластеризація даних широко використовується як групування результатів при пошуку файлів, веб-сайтів інших об'єктів, сегментації зображень, інтелектуальному аналізі даних.

Аналіз відомих розв'язань проблеми

Сьогодні найширше використовуються алгоритми ієрархічної клатеризації і алгоритми розбиття на кластери з переміщенням. До останніх належать k-means і k-medoids. Основною складністю цих алгоритмів є те, що число кластерів повинно бути задано до початку роботи алгоритмів [4]. Результати роботи і час роботи цих алгоритмів також залежать від початкового вибору кластерів. Як показано в [5] часова складність алгоритмів k-means і k-medoids пропорційна $2^{\Omega\sqrt{n}}$, де Ω – число кластерів, n – кількість об'єктів.

Побудова алгоритму

У статті пропонується алгоритм, який дає змогу визначити оптимальне число кластерів без початкових умов. Центри кластерів знаходяться за допомогою нейроподібного алгоритму.

Для заданої множини m-вимірних точок xi =(xi1, xi2, …xiN) ∈ Rm обчислюється квадратна матриця відстаней Евкліда між ними розмірністю N × N D = (Dij).

Надалі в роботі алгоритму використовуються в основному тільки ці відстані. Припустимо, що кожна точка xi є нейроном з початковою активністю Si (0)

1) для фіксованого значення порогової величини T > 0 встановлюються значення зв'язків wij між i-м і j-м нейронами

$$w_{ij} = \begin{cases} \dfrac{T^2}{D_{ij} + T^2}, якщо D_{ij} \le T \\ 0, якщо \mathrm{D_{ij}} \ge T \end{cases} \quad (3)$$

З цієї формули очевидно, що між нейронами немає зв'язків, якщо відстань між точками більша за величину T. Окрім того, wii =1, i = 1,2,…N;

2) встановлюються початкові активності для кожного нейрона

$$S_i(0) = \sum_{j=1}^{N} w_{ij} \quad (4)$$

Поступово нейрони, які перебувають на межі скупчення точок, передають свою активність нейронам, котрі розташовані в середині скупчення. Це означає, що крок за кроком нейрони, які перебувають на межі скупчень, виключаються з подальшого розгляду. Зрештою ми отримаємо ситуацію, коли залишаться лише найвіддаленіші між собою нейрони. У такому разі подальша зміна активності нейронів стає неможливою і алгоритм зупиняється;

3) змінюються активності нейронів за такою формулою

$$Si(t+1)=Si(t)+\alpha\sum_{j=1}^{N} w_{ij}\ (Si(t)-Sj(t)) \qquad (5)$$

де α – параметр, який характеризує швидкість зміни активності, $\alpha \in (0,1)$;

4) якщо в процесі зміни активності активність нейрона стає негативною Si(t)<0, тоді Si=0;

5) у результаті роботи алгоритму залишаються K нейронів, які перебувають на максимальній відстані один від одного. Точки xi, які відповідають цим нейронам, стають центрами кластерів. Всі решта точок розподіляються між кластерами за критерієм мінімуму відстані до відповідних центрів кластерів.

Кроки 1-5 повторюються для фіксованого значення порогової величини T. Якщо T = 0, тоді жоден із нейронів не може взаємодіяти з іншими нейронами (wij = 0). Всі нейрони мають одну і ту саму активність. Процес зміни активності неможливий. Отже, число кластерів починає дорівнювати числу точок N. Якщо порогова величина T є великою

$$T \geq \max (Dij)/2, \qquad (6)$$

тоді всі нейрони взаємодіють один з одним і ми отримуємо лише один кластер. Отже, істотною залишається проблема вибору значення порогової величини T.

Результати тестування алгоритму

На рис. 1 показано тестову двовимірну область, яка містить 50 точок, котрі розподілені по 5 кластерах.

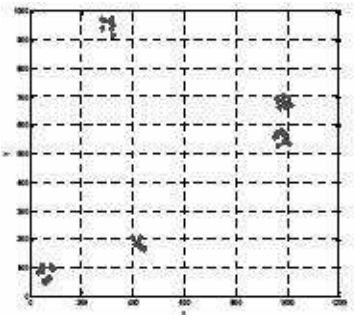

Рис. 1. Двовимірна тестова область з п'ятьма кластерами

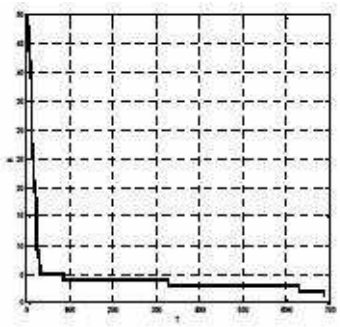

Рис. 2. Графік залежності K(T) для двовимірної тестової області з п'ятьма кластерами

На рис. 2 наведений графік залежності K(T) для заданої тестової області. З цього графіка очевидно, що на початковій стадії зміни T число кластерів змінюється дуже швидко. Із зростанням T число кластерів стабілізується на рівні K = 5 і в подальшому майже не змінюється.

На рис.3 показано тестову двовимірну область, яка містить 10 кластерів.

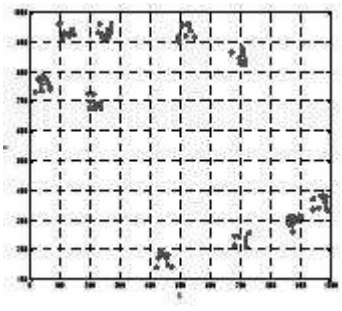

Рис. 3 Двовимірна тестова область з десятьма кластерами

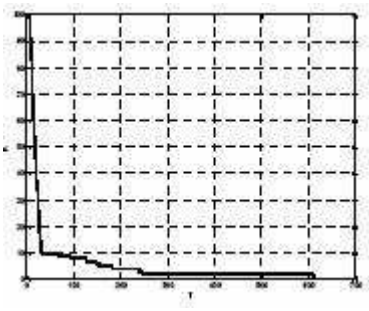

Рис. 4 Графік залежності K(T) для двовимірної тестової області з десятьма кластерами

На рис.4 показано графік залежності K(T) для заданої тестової області. Характер графічної залежності K(T) аналогічний попередньому випадку. Із зростанням T число кластерів стабілізується при K = 10. Експериментальні результати показують, що доцільно вибирати T за такою формулою:

$$T=0,1*\max(Dij)/2 \qquad (7)$$

1. Дж. Ту, Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. 2. Батыршин И.З., Хабибулин Р.Ф. Тестирование кластерных алгоритмов на инвариантность относительно нумерации объектов // Известия Академии наук. Теория и системы управления. – 1997. 3. Батыршин И.З., Хабибулин Р.Ф. Разработка алгоритмов когнитивного кластерного анализа, в кн.: Обработка текста и когнитивные технологии, вып. 3 / Под ред. В.Д. Соловьева. – Пущино, 1999. 4. Ayvazyan S., Bukhstaber V., Enyukov I., Meshalkin L. Applied statistics. Clustering and Reduction of Dimensions. Moskow.: Finansy I Statistika, 1989 5. Arthur D., Vassilvitskii S. How Slow is k-means Method? Proceedings of the 2006 Symposium on Computational Geometry (SoCG) 2006.

УДК 004.423

Білаль Раді А'Ґгель Аль-Забі, А.Б.Керницький, С.П.Ткаченко
Національний університет "Львівська політехніка",
кафедра систем автоматизованого проектування

РОЗВ'ЯЗАННЯ ЗАДАЧІ ТИПІЗАЦІЇ СХЕМ

Розглянуто питання типізації схем РЕА і встановлення їх еквівалентності.
Ключові слова – типізація, еквівалентнність, радіоелектронна схема

The problems of radioelectronic scheme typization and finding their equivalence are considered in actual paper.
Keywords – typization, equivalence, radioelectronic scheme

Вступ

Основні етапи конструкторського проектування електронних пристроїв після відпрацювання їх функціонально-логічних схем (ФЛС) наведені на рис. 1.

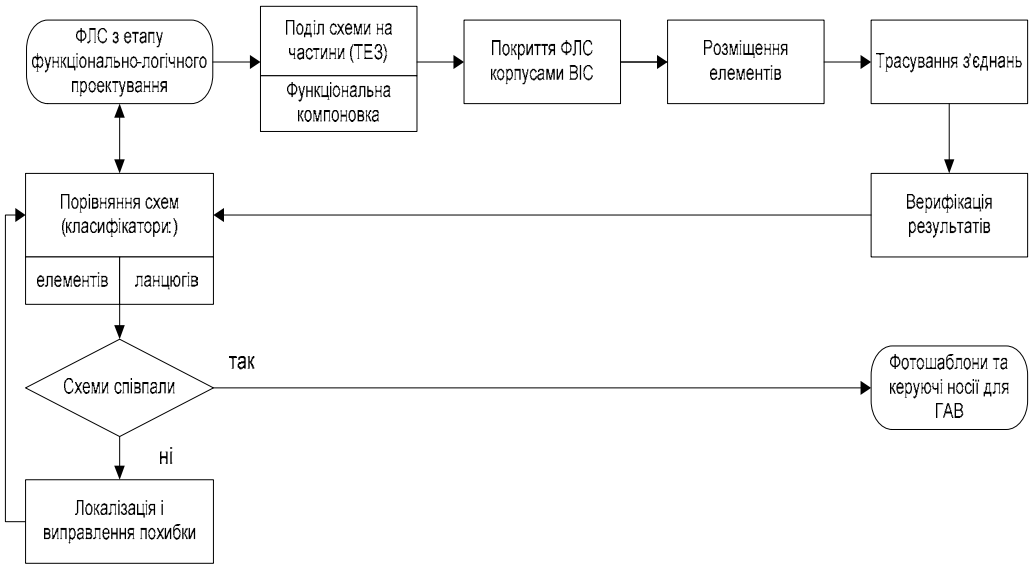

Рис. 1. Основні етапи конструкторського проектування