УДК 519.223.22:519.254

Д.І. Михайленко, П.О. Приставка

Дніпропетровський національний університет

КЛАСТЕРИЗАЦІЯ НА ОСНОВІ ОЦІНОК ФУНКЦІЇ ЩІЛЬНОСТІ БАГАТОВИМІРНИХ РОЗПОДІЛІВ

Розглянуто метод кластеризації даних на основі непараметричної оцінки функції щільності, для даних, що мають довільний закон розподілу в межах кластера.

Постановка проблеми. Задача кластеризації даних, тобто, розбиття множини на підмножини (кластери) які містять об'єкти, що мають схожі ознаки є досить актуальною в різноманітних областях науки і техніки. На сьогодні розроблено досить багато методів кластеризації даних, але вони неорієнтовані на врахування стохастичної структури множини вихідних даних. Тому на даний час є досить актуальною задача пошуку нових підходів, які б знімали питання виконання початкових умов аналізу. Для їхньої розробки необхідно мати оцінку функції щільності розподілу множини. Тоді як отримання параметричної оцінки функції щільності є досить складною задачею, то для вирішення задачі кластеризації у випадку відмінного від нормального закону розподілу в межах одного кластеру використання непараметричних оцінок функції щільності є перспективним напрямком досліджень. Таку оцінку можна отримати в результаті апроксимації локальним поліноміальним сплайном на основі B-сплайнів, близьким до інтерполяційного в середньому.

Аналіз публікацій та досліджень. Термін кластерний аналіз (уперше ввів Tryon, 1939) містить у собі набір різних алгоритмів класифікації. Загальне питання, що вирішується дослідниками в багатьох областях, полягає в тому, як організувати дані, що спостерігаються, в наочні структури, тобто розгорнуті таксономії.

У сучасних вітчизняних системах обробки статистичних даних набули поширення алгоритми кластеризації даних, такі, як: FOREL, KRAB та різного роду ієрархічні алгоритми.

Одним з найбільш уживаним з алгоритмів кластеризації є алгоритм FOREL [1] що виділяє кластери сферичної форми, кількість яких обирається користувачем. Зазначений алгоритм має ряд модифікацій які дозволяють автоматизувати вибір кількості кластерів.

Алгоритм KRAB [2] дозволяє проводити кластеризацію даних кластерів довільної форми та автоматично обирати їхню кількість. Він оснований не на врахуванні стохастичної структури множини а на розв'язанні оптимізаційних задач, що мають високу обчислювальну складність: алгоритм ґрунтується на розбитті на частини найкоротшого незамкнутого шляху (ННШ), побудова якого є досить трудомісткою задачею, і безпосередньо залежить від кількості об'єктів; кластеризація множини проводиться шляхом розбиття на частини ННШ з умови мінімізації цільової функції від багатьох змінних.

Ієрархічний метод та ряд модифікацій [3] має за основу побудову ієрархічного дерева, що демонструє структуру зв'язку множини. Своїм недоліком має високу обчислювальну складність: на кожному кроці алгоритму обчислюється матриця подібності між кластерами, яка на початку роботи алгоритму має розмірність рівну кількості елементів.

При реалізації в автоматизованих системах обробки даних зазначені алгоритми мають ряд обмежень, серед яких можна відзначити наступні: вимога нормального розподілу даних, що кластеризуються; обчислювальна складність, що суттєво зростає при збільшенні розмірності простору в якому проводиться кластеризація; використання при кластеризації допоміжних алгоритмів, що самі по собі мають високу обчислювальну складність (наприклад мінімізація цільової функції); неавтоматизоване визначення кількості кластерів; майже для всіх методів є характерним що якість кластеризації в значному ступені залежить від вибору міри відстані між кластерами.

Постановка задачі на дослідження. Нехай задана множина $\Omega_{M,N} = \left\{ x_{im,jn}, im = \overline{1,N},\ jn = \overline{1,M} \right\}$, $x_{im,jn}$ – дійсні числа, де M – кількість ознак, а N – кількість спостережень (образів).

Метою даної роботи є пошук методу, який проводить кластеризацію вихідної множини $\Omega_{M,N}$ і задовольняє наступним вимогам: не залежить від типу розподілу даних у межах кластеру, автоматично обчислює їх кількість, має низьку обчислювальну складність, може бути застосований для простору розмірності $M \geq 2$.

Викладення основного матеріалу. Вимога нормального закону розподілу припускає, що множини отримані в результаті кластеризації мають колоподібну чи овалоподібну форми. Зазвичай, така ситуація не завжди зустрічається у практичних задачах, тобто тип розподілу в межах кластеру є відмінний від нормального. У подальшому викладенні матеріалу припускається, що закон розподілу даних кожного кластера може бути відмінний від нормального.

В основі методу що пропонується, лежить побудова асимптотично точної оцінки функції щільності розподілу багатовимірної випадкової величини, що отримана в результаті апроксимації гістограми локальним поліноміальним сплайном на основі В-сплайнів, близьким до інтерполяційного у середньому [4].

Вирішення задачі кластеризації проводиться в декілька етапів: на першому етапі, в разі необхідності, проводяться координатні перетворення, які приводять до незалежності складових реалізації випадкових величин. Наступним кроком є проведення гістограмної оцінки відносних частот розбитого на класи двовимірного варіаційного ряду і оцінка за гістограмою локальних максимумів. Гістограмна оцінка апроксимується локальним поліноміальним сплайном на основі В-сплайну, близьким до інтерполяційного у середньому. За допомогою отриманої непараметричної оцінки функції щільності здійснюється уточнення розташування центроїдів. На заключному етапі проводиться кластеризація даних на основі непараметричної оцінки функції щільності з урахуванням розташування центроїдів сукупності.

Не зменшуючи загальності подальше викладення матеріалу будемо проводити для простору розмірності $M = 2$. Позначимо $X_1 = \left\{x_{l,1}, l = \overline{1,N}\right\}$, $X_2 = \left\{x_{l,2}, l = \overline{1,N}\right\}$.

На першому етапі обчислюється оцінка коефіцієнта кореляції поміж векторами спостережень X_1 та X_2, яка перевіряється на значущість [5]. У разі значущості коефіцієнта кореляції здійснюється перехід від залежних випадкових ознак X_1 і X_2 до незалежних X_1' і X_2', що пов'язано з більш високою адекватністю непараметричної оцінки щільності в межах ортогональних ознак. Виконання етапу здійснюється шляхом повороту системи координат на кут φ [6] $\forall\, g = \overline{1,N}$:

$$x'_{1g} = x_{1g}\cos\varphi + x_{2g}\sin\varphi, \qquad x'_{2g} = -x_{1g}\sin\varphi + x_{2g}\cos\varphi,$$

величина кута φ визначається із співвідношення

$$tg\,2\varphi = \frac{2\hat{r}\hat{\sigma}_1\hat{\sigma}_2}{\hat{\sigma}_1^2 - \hat{\sigma}_2^2},$$

де $\hat{r}$ – оцінка коефіцієнта кореляції поміж векторами X_1 і X_2, $\hat{\sigma}_1, \hat{\sigma}_2$ – оцінка середньоквадратичного відхилення векторів X_1 і X_2.

На другому етапі проводиться гістограмна оцінка двовимірного варіаційного ряду, в результаті виконання даного пункту отримано двовимірний варіаційний ряд $\left\{\left(x_{1i}, x_{2j}\right), n_{i,j}, p_{i,j}\ i = \overline{1,m_1}, j = \overline{1,m_2}\right\}$, за рівномірним розбиттям множини Δ_{h_1,h_2}, де m_1, m_2 – кількість елементів розбиття, h_1 і h_2 – величина кроків розбиття за напрямками X_1 і X_2 відповідно, $n_{i,j}$ – кількість елементів вихідного масиву спостережень що потрапили в межі (i,j)-го елемента розбиття Δ_{h_1,h_2}, $p_{i,j}$ – відносна частота варіанти. За варіанту ряду $\left(x_{1i}, x_{2j}\right)$ приймають центральну точку (i,j)-го елемента розбиття Δ_{h_1,h_2}:

$$x_{1i} = x_{1\min} + (i + 0.5)h_1,\ i = \overline{0, m_1 - 1},\ x_{2j} = x_{2\min} + (j + 0.5)h_2,\ j = \overline{0, m_2 - 1}.$$

На третьому етапі здійснюється пошук таких пар індексів $(mi, mj), mi = \overline{1,m_1}, mj = \overline{1,m_2}$, для яких виконується співвідношення

$$p_{mi,mj} > p_{mi+ii,mj+jj}, \text{ де } ii, jj = \overline{-1,1} \text{ та } ii, jj <> 0 \text{ одночасно.} \qquad (1)$$

Елементи варіаційного ряду з такими індексами будуть відповідати розташуванню локальних максимумів гістограми. Кінцевим результатом є масив пар індексів розташування локальних максимумів гістограми – $(mi_k, mj_k), k = \overline{1,g}$, де g – кількість локальних максимумів гістограми.

Отримана гістограма апроксимується двовимірним локальним поліноміальним сплайном $S_{r,0}$, на основі В-сплайнів, близьким до інтерполяційних у середньому

$$S_{r,0}(p,x,y) = \sum_{i\in Z}\sum_{j\in Z} p_{i,j} B_{r,h_1}(x_1 - ih_1) B_{r,h_2}(x_2 - jh_2), \qquad (2)$$

де $r = 2,3,4\ldots$, $B_{2,h}(\bullet)$ – В-сплайн r-го порядку.

У результаті отримаємо з точністю до константи h_1h_2 асимптотично-точну [4] оцінку функції щільності розподілу багатовимірної випадкової величини.

Для уточнення місця розташування центроїдів множини їх пошук, аналогічно з (1), здійснюється на $(mi_k, mj_k), k = \overline{1,g}$ – елементах розбиття Δ_{h_1,h_2}. У результаті отримаємо масив пар індексів $(si_k, sj_k), k = \overline{1,g}$, які відповідають розташуванню локальних максимумів оцінки функції щільності, що є центроїдами сукупності, що підлягає класифікації.

Головним етапом даного методу є кластеризація даних суть якої полягає в наступному.

Для всіх елементів $(x_{1l}, x_{2l})\, l = \overline{1,N}$ початкового масиву даних визначаються рівняння прямих між цим елементом і всіма локальними максимумами функції щільності. Задаючи прирощення $\Delta t_1, \Delta t_2$ отримаємо зміну координат у напрямках локальних максимумів:

$$\Delta x_{1k} = x_{1l} + (x_{1si_k} - x_{1l})\Delta t_1\,,$$

$$\Delta x_{2k} = x_{2l} + (x_{2sj_k} - x_{2l})\Delta t_2\,,\quad k = \overline{1,g}\,. \qquad (3)$$

За отриманим прирощенням відшукується напрямок максимального прирощення функції щільності з використанням непараметричної оцінки функції щільності (2) в точках (3). У результаті елемент (x_{1l}, x_{2l}) відноситься до кластеру з центроїдом в напрямку якого досягається максимальний приріст функції щільності .

Слід зауважити, що для покращення роботи методу запропоновано: проводити визначення кількості елементів розбиття за напрямками X_1 і X_2 враховуючи загальну структуру множини $\Omega_{2,N}$; після проведення кластеризації здійснити повторну кластеризацію для кожного з отриманих кластерів.

Якість роботи запропонованого методу було перевірено на результатах імітаційного моделювання, яке здійснювалось на основі алгоритмів: моделювання суміші нормальних розподілів; моделювання усіченого нормального розподілу з наперед заданою кількістю елементів у кожному.

Суттю алгоритмів моделювання даних для перевірки якості роботи методу є моделювання даних таким чином, щоб знати номер кластера до якого відноситься кожен елемент. Тоді, після застосування методу кластеризації порівнюючи отримане розташування елементів по кластерам з вихідним розташуванням елементів (еталонним) можливо оцінити похибку кластеризації. Нижче наведено опис алгоритмів для реалізації експерименту по перевірці якості роботи методу.

Алгоритм 1. Моделювання суміші розподілів:

Проводиться моделювання обраної кількості Q – двовимірних нормальних розподілів з параметрами $\vec{\Theta}_q = \{m_{1q}, m_{2q}, \sigma_{1q}, \sigma_{2q}, r_q, N_q\}$, $q = \overline{1,Q}$ де N_q – кількість елементів q-ї компоненти суміші, $\sum_{q=1}^{Q} N_q = N$. Ваговий коефіцієнт для кожної компоненти сукупності визначається з співвідношення $p_q = N_q \Big/ \sum_{k=1}^{Q} N_k$.

Алгоритм 2. Моделювання усіченого нормального розподілу з наперед заданою кількістю елементів у кожному кластері.

1. Проводиться моделювання обраної кількості E- двовимірних нормальних розподілів (допоміжних) з параметрами $\vec{\Theta}_e = \{m_{1e}, m_{2e}, \sigma_{e1}, \sigma_{e2}, r_e, N_e\}, e = \overline{1,E}$.

2. Задається величина околу d, у якому буде здійснюватися усічення даних.

4. Моделюється двовимірна нормально розподілена випадкова величина (x_1, x_2) з параметрами $\vec{\Theta} = \{m_1, m_2, \sigma_1, \sigma_2, r\ , N\ \}$ (параметри головного розподілу).

5. Якщо $\forall\ jj = \overline{1,E},\ \forall\ ii = \overline{1,N_{jj}}$ $x_1 \notin (x_{1jj,ii} - d, x_{1jj,ii} + d)$, $x_2 \notin (x_{2jj,i} - d, x_{2jj,i} + d)$, то елемент (x_1, x_2) вважається елементом кластеру. У супротивному випадку здійснюється повторна генерація величини (x_1, x_2).

6. Кроки 4 і 5 проводяться N раз.

7. Отримана множина вважається одним кластером.

8. Проводячи кроки 4–7 задану кількість Q- разів з різними пара-метрами основних розподілів отримаємо сукупність Q- усічених нормальних розподілів.

Схематично роботу алгоритму 2 наведено на графіках (Рис. 1,2).

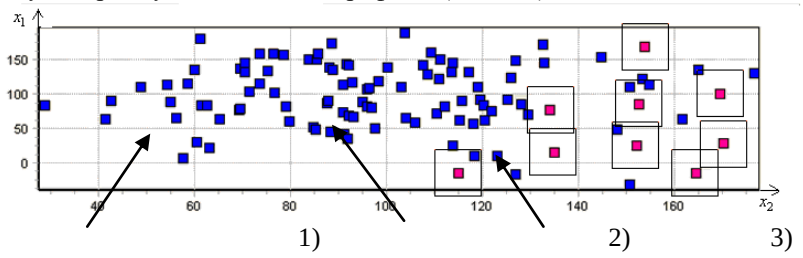


Рис. 1. Результати моделювання до застосування усічення:
1) реалізація основного розподілу, 2) реалізація допоміжного розподілу, 3) область усічення даних, діаметру *d*

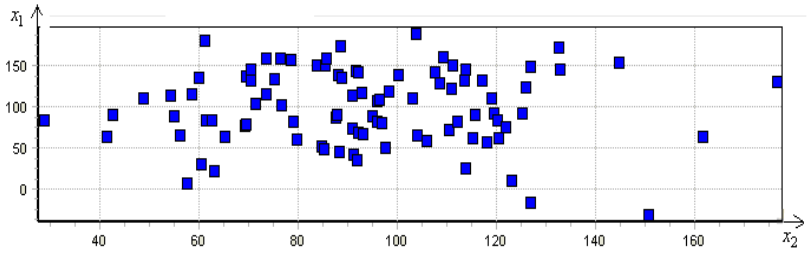


Рис. 2. Результати моделювання усіченого двовимірного розподілу

Практичні результати роботи методу отримано на основі даних, що отримані за описаними вище алгоритмами імітаційного моделювання.

Проведемо моделювання трьох неоднорідних сукупностей за алгоритмом 1 з різною кількістю елементів у кожному кластері (Табл.1).

Таблиця 1

Параметри моделювання даних за алгоритмом 1

№	m_1	m_2	σ_1	σ_1	r	N_1	N_2	N_3	N_4
1	800	800	100	100	0,5	500	50	300	400
2	1350	850	80	80	-0,2	800	80	200	400
3	1000	500	50	50	0	700	70	800	400

Кореляційне поле даних, з моделлю розподілу у вигляді суміші нормальних розподілів наведено на графіку (Рис. 3):

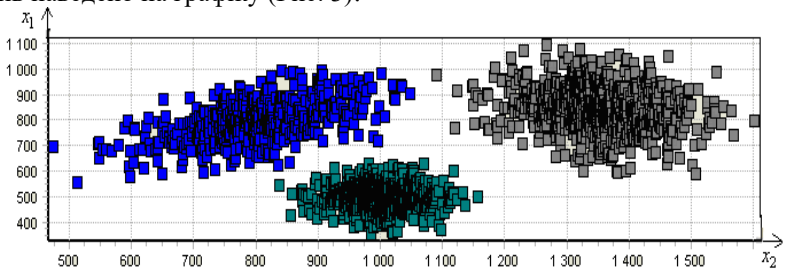
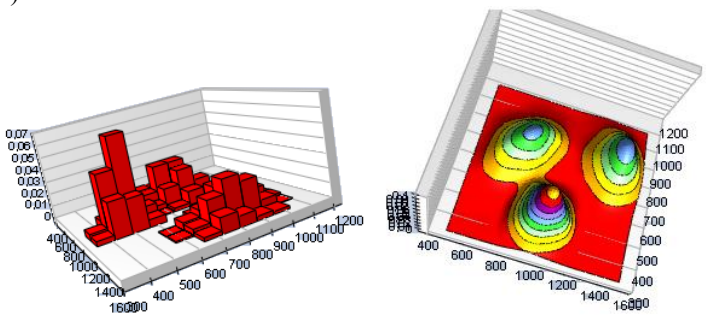


Рис. 3. Кореляційне поле даних, з моделлю розподілу у вигляді суміші нормальних

Гістограмна і непараметрична оцінки функції щільності для суміші з кількістю елементів N_3 матимуть наступний вигляд (Рис. 4):



а) б)

Рис. 4. Оцінки двовимірної неоднорідної сукупності:
а) гістограмна, б) непараметрична

Кількість елементів у кожному кластері за результатами проведення кластеризації наведено в табл.2.

Таблиця 2

Результати кластеризації множини отриманої за алгоритмом 1

Кількість елементів	Кластер 1	Кластер 2	Кластер 3	Загальна похибка
N_1	499	802	699	0,002
N_2	49	81	70	0,01
N_3	300	199	801	0,00153
N_4	400	400	400	0

Зауваження: загальна похибка обчислена як відношення кількості елементів що потрапили до класу відмінного від того в якому вони були за результатами генерації до загальної кількості елементів.

Змоделюємо за алгоритмом 2, дві неоднорідні сукупності, кожна з яких є усіченим нормальним розподілом, з наступними параметрами нормальних розподілів (Табл. 3):.

Таблиця 3

Параметри основних нормальних розподілів алгоритму 2

№	m_x	m_y	σ_x	σ_y	r	N
1	500	750	20	100	0	1000
2	500	1200	100	100	0	1000

та параметрами допоміжних нормальних розподілів (Табл. 4):

Таблиця 4

Параметри допоміжних нормальних розподілів алгоритму 2

№	m_x	m_y	σ_x	σ_y	r	N
1	500	1000	50	50	0	1000
2	500	550	10	50	0	1000

і величиною околу $d = 10$.

Кореляційне поле даних, з моделлю розподілу у вигляді усічених нормальних розподілів наведено на графіку (Рис. 5):

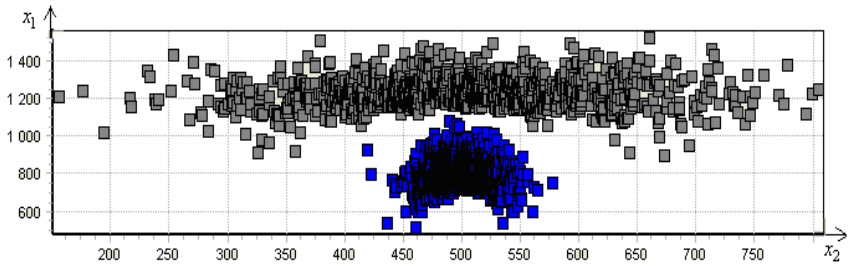
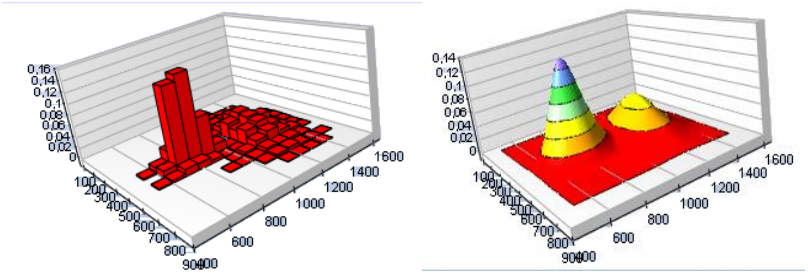


Рис. 5. Кореляційне поле даних, з моделлю розподілу у вигляді усічених нормальних розподілів

У ході виконання алгоритму кластеризації, отримано наступні гістограмну і непараметричну оцінки (Рис. 6):



а) б)

Рис. 6. Оцінки двовимірної неоднорідної сукупності:
а) гістограмна, б) непараметрична

Кількість елементів у кожному кластері за результатами проведення кластеризації наведено в табл. 5:

Таблиця 5

Результати кластеризації множини отриманої за алгоритмом 2

Кластер	Фактична кількість	Відхилення	Похибка
1	997	3	0,003
2	1003	3	0,003

Висновки. У роботі запропоновано метод та обчислювальну технологію кластеризації даних, яка не залежить від типу розподілу даних, проводить автоматичне обчислення кількості кластерів, має низьку обчислювальну складність, може бути застосований до простору розмірності $M \geq 2$. Результати роботи методу протестовано на підставі методів імітаційного моделювання серед яких окремий інтерес представляє запропонований алгоритм моделювання суміші усічених нормальних розподілів.

Подальшим напрямком є дослідження інформаційної технології кластеризації даних для простору розмірності $M > 2$, розробка методу відтворення оцінок параметрів багатовимірних сумішей нормальних розподілів і методів класифікації даних на основі непараметричної оцінки функції щільності.

Результати досліджень отримані при виконанні теми за проектом Ф7/382-2001 Державного фонду

фундаментальних досліджень.

Бібліографічні посилання

1. **Айвазян С.А.** Классификация многомерных наблюдений / С.А. Айвазян, З.И. Бежаева, О.В. Староверов. – М., 1974. – 240 с.
2. **Загоруйко Н.Г.** Методы обнаружения закономерностей. Серия «Математика, кибернетика», №11. –, М., 1981. – Х с.
3. **Жамбю. М.** Иерархический кластер-анализ и соответствия. – М., 1988. –342 с.
4. **Приставка П.О.** Поліноміальні сплайни при обробці даних. – Д., 2004. – 236 с.
5. **Бабак В.П.** Статистична обробка даних / В.П. Бабак, А.Я. Білецький, О.П. Приставка. – К., 2001. – 388 с.
6. **Коваленко И.Н.** Теория вероятностей / И.Н. Коваленко, Б.В. Гнеденко. – К., 1990. – 328 с.

Надійшла до редколегії 11.05.06