Кібернетика та системний аналіз

УДК 004.318 DOI: 10.30748/zhups.2018.56.11

В.П. Бурдаєв

Харківський національний економічний університет ім. С. Кузнеця, Харків

СТРУКТУРА БАЗИ ЗНАНЬ ДЛЯ ВИБОРУ АЛГОРИТМУ КЛАСТЕРИЗАЦІЇ ДАНИХ

У статті обґрунтована актуальність завдання вибору алгоритму кластеризації об'єктів, на основі знань експертів. Проблема кластеризації полягає в тому, що для кожного конкретного типу даних, структури розташування об'єктів в просторі ознак треба або правильно підібрати відомий алгоритм, або його адаптувати або розробити новий. Вивчається математична модель кластерного аналізу. Адаптовано алгоритм динамічних згущень для розмитої класифікації різнорідних даних. Викладається стратегія проведення кластеризації в системі "КАРКАС". Розглянуто приклади правил і фреймів бази знань для вибору алгоритму кластерного аналізу.

Ключові слова: *кластерний аналіз, база знань, експертна система.*

Вступ

Важливим моментом в кластерному аналізі вважається вибір метрики (міри близькості об'єктів), від якого вирішальним чином залежить остаточний варіант розбиття об'єктів на групи при заданому алгоритмі розбиття [1–4].

Іншою важливою величиною в кластерному аналізі є відстань між кластерами об'єктів. Вибір тієї чи іншої міри відстані між кластерами залежить від геометричних фігур, які утворюють об'єкти в просторі ознак. Наприклад, застосування відстані "найближчого сусіда" має хороші результати кластеризації, коли об'єкти в просторі ознак утворюють цепочну структуру. Відстань "далекого сусіда" застосовується, коли об'єкти утворюють кулясті хмари. У разі, коли об'єкти утворюють еліпсоїди, то рекомендується використовувати відстані між їх центрами тяжіння.

Алгоритми кластерного аналізу відрізняються великою різноманітністю. Це можуть бути, наприклад, алгоритми, що реалізують повний перебір об'єктів або здійснюють випадкові розбиття множини об'єктів. У той же час більшість таких алгоритмів складається з двох етапів. На першому етапі задається початкова (випадкове) розбиття множини об'єктів на кластери і визначається функціонал якості розбиття. На другому етапі, об'єкти переносяться з кластера в кластер до тих пір, поки значення функціоналу якості розбиття не перестане поліпшуватися.

Проблема кластеризації полягає в тому, що для кожного конкретного типу даних, структури розташування об'єктів в просторі ознак, треба або правильно підібрати відомий алгоритм, або його адаптувати або розробити новий. Для вирішення цієї проблеми широко застосовують знання експертів [5–6].

Метою даної роботи є дослідження структури бази знань предметної області для кластерного аналізу. Дана робота є розвитком досліджень [7–10].

Постановка задачі. Розробити ефективну структуру бази знань для вибору алгоритму кластеризації об'єктів.

Основна частина

Нехай $X = \{x_1, x_2, ..., x_n\}$ – множина об'єктів, кожен з яких описується одним і тим же набором з P ознак. Тоді X можна розглядати як підмножину простору ознак $P = P_1 \times P_2 \times ... \times P_p$, де P_i – множина значень i-ї ознаки і знак $\times$ – прямий добуток.

Завдання кластер-аналізу полягає в тому, щоб отримати розбиття множини X на систему підмножин $\{S_1, S_2, ..., S_k\}$ таку, що $X = S_1 \cup S_2 \cup, ..., \cup S_k$, $i \neq j$, $i, j = 1, 2, ..., k$, і задовольняє певному критерію якості розбиття, наприклад, мінімуму внутрішньогрупової суми квадратів відхилень (приклад, функціоналу якості розбиття).

Будь-яке розбиття множини X, взаємнооднозначно визначає ставлення еквівалентності $R \subset X \times X$. Тоді фактормножина X/R задає відповідне розбиття множини X, а канонічна проекція $Pr_1 : X \to X/R$ – розподіл об'єктів по кластерам.

Нехай k – потужність фактормножини X/R і ототожним його з множина $N = \{1, 2, ..., k\}$ – номерів класів еквівалентності, тоді для будь-якого $t \in N$ кластер S_t визначається наступним чином: $S_t : q \circ Pr_1(X) = t$, де q – ізоморфізм X/R на N.

Нехай Θ – множина всіх відносин еквівалентності множини X. Ясно, що не всяке відношення еквівалентності являє собою цінність для вирішення завдання кластер-аналізу. Тому для вибору фактормножини, що відображає можливий природний розподіл об'єктів на кластери, вводиться функціонал F на Θ. Тоді число $F(\Theta)$ можна розглядати як

числову характеристику розбиття. Цей функціонал називають або функціоналом якості розбиття, або цільовою функцією, або критерієм якості розбиття.

Отже, математична модель кластер-аналізу – це трійка об'єктів (P, F, Θ).

Процес кластеризації можна розглядати як ітеративне породження фактормножин по заданому відношенню еквівалентності. Дійсно, в силу транзитивності факторизації, по індукції отримуємо наступний ланцюжок незростаючих по потужності фактормножин: $F_1, F_2, ..., F_h, ...$, де $F_1 = X / R$, $F_h = F_{h-1} / R$. Стабілізація цього ланцюжка або її обрив на певному кроці здійснюється за допомогою функціоналу F. Наприклад, нехай при агломеративному ієрархічному угрупованні об'єктів відношення еквівалентності між ними визначається за принципом "найближчого сусіда" і "K" – число кластерів. Тоді для того, щоб отримати розбиття на "K" кластерів, треба побудувати ланцюжок з "N – K" факторможин, де останнє дає шукане розбиття. В інших алгоритмах кластеризації типу "К-внутрішньогрупових середніх", ланцюжок факторможин стабілізується шляхом коригування центрів кластерів, а в алгоритмі "ISODATA" за допомогою евристичних параметрів здійснюється її коливання, тобто поряд з об'єднанням кластерів відбувається розщеплення їх, і остаточний варіант вибору розбиття залежить від дослідника. Таким чином, для того, щоб вирішити завдання кластерного аналізу, треба забезпечити простір ознак P метрикою і підібрати відношення еквівалентності з Θ, яке давало б, наприклад, екстремум деякому функціоналу F.

Наявність різних альтернатив в реалізації математичної моделі кластерного аналізу якраз і передбачає використання знань експертів по класифікації для отримання природного розбиття об'єктів на кластери. Останнє твердження стимулює розробки зі створення експертних систем, які могли б їх використати при кластеризації.

Алгоритм динамічних згущень для розмитої класифікації різнорідних даних. Нехай $X = \{x_1, x_2, ..., x_n\}$ – кінцева множина об'єктів. Кожен об'єкт x_i описується кінцевим набором з p ознак, тобто $x_i = (x_{i1}, x_{i2}, ..., x_{in})$. Тоді X можна розглядати як підмножину множини $P = P_1 \times P_2 \times ... \times P_p$, де P_j – множина значень j-ї ознаки. Якщо ознака кількісна, то $P_j = R$ – множина дійсних чисел. Якщо ознака порядкова або номінальна, то P_j – множина міток довільної природи.

Система множин $S = (S_1, S_2, ..., S_k)$ називається k-покриттям множина X, якщо виконані наступні умови:

1. $S_i \neq \varnothing$ для $i = 1, 2, ..., k$;

2. $X = \bigcup_{i=1}^{k} S_i$.

Задамо на множині X структуру представництв, поставивши йому у відповідність множина представництв Z і відображення $D' : X \times Z \to R^+$ [1]. Нехай 2^X – множина всіх підмножин в X. Розглянемо відображення $D : 2^X \times Z \to R^+$, таке, що

$$D(\{x\}, z) = D'(x, z) \qquad (x \in X, z \in Z).$$

Припустимо, що виконані наступні умови:

1. $\min D(A, \bullet)$ існує і єдиний $(A \in 2^X)$;

2. $D(A, z) = \sum D(\alpha, z)$; $(\alpha \in A, A \in 2^X, z \in Z)$;

3. $Z = P$.

Оскільки ознаки різнотипні і можуть мати різну інформативність, то міру подібності D' між об'єктом $x \in X$ і представником $z \in Z$ локально пристосуємо до структури множини об'єктів X і задамо формулою

$$D'(x, z) = \sum_{l=1}^{p} \alpha_l d^l(Pr_l x, Pr_l z) / \sum_{l=1}^{p} \alpha_l,$$

де $Pr_l : P \to P_l$ – ортогональний проектор; α_l – коефіцієнт інформативності l-ї ознаки, $0 \le \alpha_l \le 1$; відображення d^l визначається в залежності від типу ознаки.

Нехай l-а ознака кількісна. Тоді

$$d^l : Pr_l(X) \times Pr_l(Z) \to R^+;$$
$$d^l(Pr_l x, Pr_l z) = | Pr_l x - Pr_l z | / \qquad (1)$$
$$/ | \max Pr_l(X) - \min Pr_l(X) |.$$

У разі, якщо $\max Pr_l(X) = \min Pr_l(X)$, то $d^l(Pr_l x, Pr_l z) = 0$. Нехай l-а ознака порядкова. Розглянемо міру відмінності

$$\tilde{d} : Pr_l(X) \times Pr_l(X) \times Pr_l(X) \to R^+$$

у відносинах $Pr_l x$ і $Pr_l z$ до деякого елементу $y \in Pr_l(X)$, певну наступним чином: $\tilde{d}(Pr_l x, Pr_l z, y) = 0$, якщо $(Pr_l x - y)(Pr_l z - y) \ge 0$ або $Pr_l x = Pr_l z = y$.

$\tilde{d}(Pr_l x, Pr_l z, y) = 1$, якщо $(Pr_l x - y)(Pr_l z - y) \le 0$.

$\tilde{d}(Pr_l x, Pr_l z, y) = 0{,}5$, якщо $\begin{cases} Pr_l x = y, & \text{а} \quad Pr_l z \neq y; \\ Pr_l x \neq y, & \text{а} \quad Pr_l z = y. \end{cases}$

Тоді

$$d^l(Pr_l x, Pr_l z) =$$
$$= (1/(n-1)) \sum_{y \in Pr_l(X)} \tilde{d}(Pr_l x, Pr_l z, y). \qquad (2)$$

Для номінальної l-ї ознаки міра відмінності $\tilde{d}$ визначається наступним чином:

$$\tilde{d}\left(Pr_l x, Pr_l z, y\right) = 0, \text{ якщо } \begin{cases} Pr_l x = Pr_l z = y; \\ Pr_l x \neq y \text{ та } Pr_l z \neq y. \end{cases}$$

$$\tilde{d}\left(Pr_l x, Pr_l z, y\right) = 1, \text{ якщо } \begin{cases} Pr_l x = y; \quad Pr_l z \neq y; \\ Pr_l x \neq y; \quad Pr_l z = y. \end{cases}$$

Тоді $$Z_j^{(N)} \neq Z_j^{(N+1)} \qquad (3)$$

Функції належності для множин k-покриття $S = \left(S_1, S_2, ..., S_k\right)$ множина X визначимо наступним чином:

$$W_j : X \to [0,1]; \quad \left(j = 1,2,...,k\right);$$

$$w_{ij} \geq 0; \quad \sum_{j=1}^{k} w_{ij} = 1; \quad \left(i = 1,2,...,n\right),$$

де $w_{ij} = W_j\left(x_i\right)$ визначають ступінь належності об'єкта x_i до множина S_j.

Нехай $D\left(S_j, Z_j\right) = \sum_{x_i \in S_j} w_{ij}^m D'\left(x_i, z_j\right), \quad \left(i = 1,2,...,n\right),$

де $z_j \in Z$ – представник елемента покриття S_j; $j = 1,2,...,k$; $m > 2$ – ціла вагова константа.

Нехай $S^\Delta = \{S\}$ – множина k-покриттів X і $Z^\Delta = Z^k$. Функціоналом якості покриття назвемо перетворення

$$F : S^\Delta \times Z^\Delta \to R^+;$$

$$F\left(S, Z\right) = \sum_{j=1}^{k} D\left(S_j, Z_j\right) \left((S, Z) \in S^\Delta \times Z^\Delta\right).$$

Оптимізаційна задача формулюється так: знайти k-покриття S^* і представництво Z^* які доставляють мінімум функціоналу F, тобто

$$F\left(S, Z\right) \to \min. \qquad (4)$$

Спочатку вирішимо завдання по знаходженню мінімуму по змінній Z (при фіксованій S). Визначимо функцію представництва [1] як перетворення $g : S^\Delta \to Z^\Delta$, таке, що $g\left(S\right) = Z^*$ при $S = \left(S_1, S_2, ..., S_k\right)$ і $Z^* = \left(Z_1^*, Z_2^*, ..., Z_k^*\right)$, де Z_j^* мінімізує $D\left(S_j, Z_j\right)$.

$$D\left(S_j, Z_j\right) = \sum_{x_i \in S_j} w_{ij}^m D'\left(x_i, z_j\right) = \sum_{i=1}^{n} w_{ij}^m D'\left(x_i, z_j\right) =$$

$$= \sum_{i=1}^{n} w_{ij}^m \sum_{l=1}^{p} \alpha_l d^l\left(Pr_l x_i, Pr_l z_j\right) / \sum_{l=1}^{p} \alpha_l =$$

$$= \left(\sum_{l=1}^{p} \alpha_l\right)^{-1} \sum_{l=1}^{p} \alpha_l \sum_{i=1}^{n} w_{ij}^m d^l\left(Pr_l x_i, Pr_l z_j\right) \left(z_j \in Z\right).$$

Отже, для $j = 1,2,...,k$ і $l = 1,2,...,p$ необхідно мінімізувати

$$\sigma\left(Pr_l z_j\right) = \sum_{i=1}^{n} w_{ij}^m d^l\left(Pr_l x_i, Pr_l z_j\right) \quad \text{по} \quad \left(z_j \in Z\right). \qquad (5)$$

Для кожного типу ознаки знайдемо явний вид функції σ. Для l-ї кількісної ознаки відповідно до формули (1) маємо:

$$\sigma\left(Pr_l z_j\right) = \sum_{i=1}^{n} w_{ij}^m \left| Pr_l x_i - Pr_l z_j \right| /$$

$$/ \left| \max Pr_l\left(X\right) - \min Pr_l\left(X\right) \right|.$$

Функція $\sigma\left(Pr_l z_j\right)$ – опукла, кусочно-лінійна функція однієї змінної на R і мінімум її може досягатися в одній з точок множини $A = \left\{Pr_l x_i \mid i = 1,2,...,n\right\}$, тобто мінімум існує, але може бути неєдиним.

Нехай l-а ознака порядкова, має γ можливих значень $G_1, G_2, ..., G_\gamma$, занумерованих в порядку зростання, u_i – кількість об'єктів в X, що мають по l-й ознаці значення G_i, $i = 1,2,...,\gamma$.

Допустима множина рішень для задачі (5) є кінцева множина $\left\{G_1, G_2, ..., G_\gamma\right\}$, а значення функції σ в точці G_t, $t = 1,2,...,\gamma$ можна обчислити, згідно з формулою (2), слідуючим чином:

$$A_1 = \left\{\sum_{i=1}^{t-1} u_i \left(\sum_{j=1}^{i-1} r_j + 0,5 r_i\right)\right\};$$

$$A_2 = \left\{0,5 u_t \left(\sum_{i=1}^{t-1} r_i + \sum_{i=t+1}^{\gamma} r_i\right)\right\};$$

$$A_3 = \sum_{i=t+1}^{\gamma} u_i \left(\sum_{j=i+1}^{\gamma} r_j + 0,5 r_i\right);$$

$$\sigma\left(G_t\right) = (1/(n-1))\left\{A_1 + A_2 + A_3\right\},$$

де $r_j = \sum_{q \in Q_j} w_{qj}^m$, $Q_j = \left\{i \in \left\{1,2,...,n \mid Pr_l x_i = G_j\right\}\right\}$ – множина номерів об'єктів з X, що приймають значення G_j по l-й ознаці, $j = 1,2,...,\gamma$.

Для номінального ознаки відповідно до формули (3) маємо:

$$\sigma\left(G_t\right) = \sum_{i=1}^{t-1} \left(u_i + u_t\right) r_i + \sum_{i=t+1}^{\gamma} \left(u_i + u_t\right) r_i$$

для $t = 1,2,...,\gamma$.

Таким чином, мінімум завдання (5) по змінній Z існує, але може бути неєдиним, і є формули, які дозволяють його знайти.

Розглянемо другу задачу по визначенню мінімуму по змінної S (при фіксованій Z). Визначимо функцію призначення [1] як перетворення $f : Z^\Delta \to S^\Delta$, таке, що $f\left(Z\right) = S^*$, де $Z = \left(z_1, z_2, ..., z_k\right)$, $S^* = \left(S_1^*, S_2^*, ..., S_k^*\right)$.

Для $j = 1,2,...,k$ множини $S_j^* = \left\{x_i \mid w_{ij}^* > 0\right\}$, де вектора $w_i^* = \left(w_{i1}^*, w_{i2}^*, ..., w_{ik}^*\right)$, $i = 1,2,...,n$, підбираються так, щоб

$$\sum_{j=1}^{k} w_{ij}^{m} D'\left(x_i, z_j\right) \to \min, \qquad (6)$$

при обмеженнях

$$w_{ij} > 0;\ \sum_{i=1}^{n} w_{ij} > 0;\ \sum_{j=1}^{k} w_{ij} = 1. \qquad (7)$$

Завдання (6–7) є задачею лінійного програмування від змінної w_i, тому мінімум існує і досягається в одній з вершин багатогранника рішень (7). Знайдемо рішення. Дотримуючись методу множників Лагранжа, розглянемо функціонал

$$L\left(w_i, \lambda\right) = \sum_{j=1}^{k} w_{ij}^{m} D'\left(xi, zj\right) - \lambda\left(\sum_{j=1}^{k} w_{ij} - 1\right).$$

Маємо

$$\frac{\partial L}{\partial w_{ij}} = m\left(w_{ij}\right)^{m-1} D'\left(x_i, z_j\right) - \lambda = 0,$$

$$j = 1, 2, ..., k.$$

Отже,

$$w_{ij} = \left[\frac{m}{\lambda} D'\left(x_i, z_j\right)\right]^{m-1};\ j = 1, 2, ..., k,$$

де $\sum_{j=1}^{k}\left[\frac{m}{\lambda} D'\left(x_i, z_j\right)\right]^{m-1} = 1.$

Таким чином, множина S_j^* визначається як множина об'єктів x_i і з функцією належності w_j^*, значення якої обчислюються за формулою

$$w_{ij}^{*} = \left[D'\left(x_i, z_j\right)\right]^{m-1}\left[\sum_{q=1}^{k} D'\left(x_i, z_q\right)\right]^{-(m-1)} \qquad (8)$$

для будь-якого x_i з умовою $D'\left(x_i, z_q\right) > 0$, $q = 1, 2, ..., k$. Якщо існує об'єкт x_i, такий що $D'\left(x_i, z_q\right) = 0$ для деякого q, тоді $w_{iq}^{*} = 1$, $w_{ij}^{*} = 0$, $j \neq q$, $j = 1, 2, ..., k$.

Тепер опишемо ітераційний алгоритм, який дає рішення задачі (4). Початкове k-покриття $S^{(0)}$ задамо довільними функціями належності

$$W_j : X \to [0,1], \quad j = 1, 2, ..., k;$$

$$w_{ij} > 0; \quad \sum_{i=1}^{n} w_{ij} > 0; \quad \sum_{j=1}^{k} w_{ij} = 1;$$

$$\left(1 \le i \le n, 1 \le j \le k\right).$$

Рух алгоритму визначимо наступним чином:

$$v_n = \left(S^{(n)}, Z^{(n)}\right); \quad Z^{(n)} = g\left(S^{(n-1)}\right);$$

$$\text{та} \quad S^{(n)} = f\left(Z^{(n)}\right); \quad u_n = F\left(v_n\right),$$

тобто процес побудови оптимального рішення задачі (4) полягає в ітеративному застосуванні функцій g і f. При цьому, на кожному кроці ітерації алгоритму зменшується значення функціоналу F. В силу, що не один мінімум для функції представництва, пропонується наступний спосіб знаходження мінімуму. Мінімум на кроці алгоритму $N+1$ вибирається рівним мінімуму на кроці N. Якщо це не можливо виконати, то мінімум вибирається довільним чином.

У реальних застосуваннях в процесі прагнення до глобального мінімуму можуть зустрічатися локальні мінімуми, викликані або невдалим вибором мінімуму функції представництва, або вибором початкового k-покриття $S^{(0)}$.

Сходимість алгоритму випливає з пропозиції, доказ якого проводиться аналогічно як в [1].

Пропозиція. Послідовності u_n і v_n сходяться і стабілізуються одночасно за кінцеве число кроків.

Доведення. Покажемо, що

$$u_{n-1} \ge F\left(S^{(n-1)}, Z^{(n)}\right) \ge u_n.$$

Перша нерівність справедлива, оскільки для $j = 1, 2, ..., k$

$$D\left(S_j^{(n-1)}, Z_j^{(n-1)}\right) \ge D\left(S_j^{(n-1)}, Z_j^{(n)}\right) \qquad (9)$$

по визначенню g. Підсумовуючи по j від 1 до k, отримуємо:

$$u_{n-1} = F\left(S^{(n-1)}, Z^{(n-1)}\right) \ge F\left(S^{(n-1)}, Z^{(n)}\right). \qquad (10)$$

Друга нерівність справедлива, оскільки для $i = 1, 2, ..., n$

$$\sum_{j=1}^{k}\left(w_{ij}^{(n-1)}\right)^{m} D'\left(x_i, z_j^{(n)}\right) \ge \sum_{j=1}^{k}\left(w_{ij}^{(n)}\right)^{m} D'\left(x_i, z_j^{(n)}\right)$$

вже за визначенням f. Підсумовуючи по i від 1 до n, маємо:

$$F\left(S^{(n-1)}, Z^{(n)}\right) \ge F\left(S^{(n)}, Z^{(n)}\right) = u_n. \qquad (11)$$

Послідовність u_n убуває і обмежена знизу, отже, сходиться. Будь-яка послідовність, що сходиться на кінцевій множині, стабілізується, тобто починаючи з деякого кроку N, $u_n = u_N$ для $n > N$.

Доведемо сходимість послідовності v_n.

Нехай послідовність u_n сходиться на N-му кроці алгоритму, тобто $u_N = u_{N+1}$. Тоді $F\left(v_N\right) = F\left(v_{N+1}\right)$. Покажемо, що $v_n = v_N$ для всякого $n > N$. Справді, $F\left(v_N\right) = F\left(v_{N+1}\right)$ або $F\left(S^{(N)}, Z^{(N)}\right) = F\left(S^{(N+1)}, Z^{(N+1)}\right)$ і $g\left(S^{(N)}\right) = Z^{(N+1)}$.

Звідси випливає, що $Z^{(N)} = Z^{(N+1)}$. Дійсно, припустимо противне, тобто $Z^{(N)} \neq Z^{(N+1)}$. Це означає, що існує індекс j, $j \in \{1, 2, ..., k\}$ такий, що $Z_j^{(N)} \neq Z_j^{(N+1)}$. Отже, нерівності (9–10) перетворюються в строгі, тобто маємо:

$$u_N = F\left(S^{(N)}, Z^{(N)}\right) > F\left(S^{(N)}, Z^{(N+1)}\right).$$

З іншого боку, відповідно до нерівності (11), маємо

$$F\left(S^{(N)}, Z^{(N+1)}\right) \geq F\left(S^{(N+1)}, Z^{(N+1)}\right) = u_{N+1}.$$

Отже, $u_N > u_{N+1}$. Це суперечить припущенню, що $u_N = u_{N+1}$.

Таким чином, $Z^{(N)} = Z^{(N+1)}$, звідки $f\left(Z^{(N)}\right) = f\left(Z^{(N+1)}\right)$ за визначенням f, отже, $S^{(N)} = S^{(N+1)}$, звідки $v_N = v_{N+1}$. Таке міркування можна провести для будь-якого $n \geq N$, тобто $v_n = v_N$ $\left(n \geq N\right)$. Пропозиція доведена.

Розглянемо окремий випадок k-покриття $S = \left(S_1, S_2, ..., S_k\right)$ множина X. Нехай виконано ще одну додаткову умову:

$$S_i \bigcap S_j = \varnothing, \quad i \neq j, \quad i, j = 1, 2, ..., k. \qquad (12)$$

Покриття, що задовільняють умові (12), називаються розбивками, які відіграють важливу роль в проблемі класифікації. Елементи k-розбиття називаються кластерами.

Покажемо, як можна з k-покриття S множина X виділити k-розбиття. Нехай $x_i \in X$ належить кільком елементам покриття $S = \left(S_1, S_2, ..., S_k\right)$, тобто $x_i \in \bigcup S_j$, $j \in Q \subseteq \left\{1, 2, ..., k\right\}$ і $w_{ir}^* = \max w_{ij}$, $j \in Q$. Тоді $x_i \in S_r^*$. Якщо таких $r \in Q$ кілька, то можна взяти будь-який, наприклад, найменше.

Якщо об'єкт $x_i \in X$ належить тільки одному елементу покриття S_j, тобто $w_{ij} = 1$, $w_{iq} = 0$, $q \neq j$, $j = 1, 2, ..., k$. Тоді об'єкт x_i буде визначати ядро для кластера S_r^*.

Таким чином, будується k-розбиття $S^* = \left(S_1^*, S_2^*, ..., S_k^*\right)$.

За отриманим k-покриттям можна зробити ряд евристичних рекомендацій щодо взаємного розташування об'єктів в просторі ознак.

Наприклад, можна виділити "прикордонні" об'єкти для кластерів, тобто такі об'єкти $x_i \in X$, для яких значення функцій приналежності $w_{ij} = 1/k$, $j = 1, 2, ..., k$.

Якщо число k (число кластерів) велике (більше, ніж 100), то можна вважати, що "прикордонні" об'єкти вироджуються в "аномальні", мало впливають на процес кластеризації.

Аналогічно будується ітераційний алгоритм близький до алгоритму "ISODATA". Тут додані допоміжні евристичні процедури по розщепленню і об'єднанню кластерів на основі обчислення їх статистик [11–14].

Для вирішення завдання кластерного аналізу за допомогою системи "КАРКАС" [10] рекомендується виконати ряд операцій:

1. Матрицю "об'єкт-ознака" в форматі Excel завантажити за допомогою модуля системи кластеризація.

2. Ввести параметри кластеризації і виконати алгоритм "К-внутрішньогрупових середніх" або "ISODATA".

3. Аргументувати результати кластеризації: на екран дисплея виводиться таблиця відстаней між центрами кластерів, таблиця дисперсій для отримання уявлення про відносне розташування образів в середині кластера і інше.

Така стратегія кластеризації дозволяє експерту отримати додаткову інформацію про кількість, форму і компактність кластерів, про кількість центрів кластерів і їх координатах, про відстані між кластерами і про розміри "аномальних" кластерів.

На основі отриманих результатів експерт складає правила для БЗ [15] за допомогою візуального редактора БЗ системи "КАРКАС".

У режимі консультації з користувачем система вибирає необхідні правила для формування алгоритму кластеризації.

Кожен метод або алгоритм кластеризації може застосовуватися в певній ситуації, яка описується різною інформацією: поганим або хорошим розташуванням об'єктів у просторі ознак, щільністю кластерів і тому подібному.

Робота машини виведення системи полягає в тому, що вона аналізує ієрархічну функціональну систему побудовану системою "КАРКАС" для вибору відповідних параметрів і алгоритмів кластеризації. Наведемо кілька правил і фреймів.

Правіло_1. А #.
ЯКЩО
А Ознаки = Кількісні
ТО
Метрика = Евклидова.
Правіло_2. А #.
ЯКЩО
А Ознаки = Порядкові
ТО
Метрика = інфімум.
Фрейм_16. Ім'я слота | Тип слота | спадкування
вибір | заміщення |
метрика | Не використовується |
число | ні |
потужність | заміщення |
Ступінь належності | заміщення |
Цільовий слот.
кластеризація | "Нечітке динамічне згущення".
Фрейм для вибору процедури нечіткого динамічного згущення.
Фрейм_18. Ім'я слота | Тип слота | спадкування
стратегія | невідома |
число | ні |

аномал | заміщення |
центр | заміщення |
Цільовий слот.
кластеризація | "К-внутрішньогрупових середніх".
Фрейм_19.
Ім'я слота | Тип слота | спадкування
стратегія | невідома |
число | введення |
аномал | заміщення |
центр | заміщення |
дисперсія | заміщення |
відстань | заміщення |
ітерація | заміщення |
розщеплення | заміщення |
Цільовий слот.
кластеризація | "ISODATA".

Вибір відповідного правила або фрейма здійснюється за допомогою ієрархічної функціональної системи [8].

Висновок

Результатом роботи була розробка структури бази знань для вибору алгоритму кластеризації. Адаптовано алгоритм динамічних згущень для розмитої класифікації різнорідних даних. Алгоритм дозволяє проводити кластеризацію різнотипних об'єктів, з урахуванням функції належності. Цей алгоритм, реалізований у вигляді програмного модуля, може бути викликаний за допомогою фрейму бази знань.

Подальшим напрямком даного дослідження може стати розроблення інших алгоритмів для бази знань кластерного аналізу.

В ході виконання роботи створено модуль кластеризації, який імплементований в систему "КАРКАС" (електронний ресурс системи https://it-karkas.com.ua).

Список літератури

1. Дидэ Э. Методы анализа данных /Э. Дидэ. – М.: Финансы и статистика, 1985. – 357с.
2. Жамбю М. Иерархический кластер-анализ и соответствия / М. Жамбю. – М.: Финансы и статистика, 1988. – 342 с.
3. Черноруцкий И.Г. Методы принятия решений / И.Г. Черноруцкий. – СПб.: БХВ-Петербург, 2005. – 416 с.
4. Технологии анализа данных / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко и др. – СПб.: БХВ-Петербург, 2008. – 384 с.
5. Жданов А.А. Автономный искусственный интеллект /А.А. Жданов. – М.: БИНОМ. Лаборатория знаний, 2008. – 359 с.
6. Башмаков А.И. Интеллектуальные информационные технологии / А.И. Башмаков, И.А. Башмаков. – М.: МГТУ им. Н.Э. Баумана, 2005. – 304 с.
7. Бурдаев В.П. Сложность динамических систем: моногр. / В.П. Бурдаев. – Saarbrucken: LAP LAMBERT Academic Publishing, 2015. – 140с.
8. Burdaev V.P. About one concept of constructing a temporal knowledge base / V.P. Burdaev // Fundamental and Applied Studies in the Pacific and Atlantic Oceans Countries: in Proc. of the 1st International Congress, Tokyo University Press, 2014, p. 272-276.
9. Бурдаев В.П. Системи навчання з елементами штучного інтелекту: моногр. / В.П. Бурдаєв. – Х.: Вид. ХНЕУ, 2009. – 392 с.
10. Бурдаев В.П. Моделі баз знань: моногр. / В.П. Бурдаєв. – Х.: Вид. ХНЕУ, 2010. – 300 с.
11. Hae-Sang Park, Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications. 36. – 2009. – P. 3336-334.
12. Khan S.S., & Ahmad A. Cluster center initialization algorithm for K-means clustering // Pattern Recognition Letters, 25. – 2004. – P. 1293-1302.
13. Zhang Q., & Couloigner I. A new and efficient k-medoid algorithm for spatial clustering // Lecture Notes in Computer Science, 3482. – 2005. – P. 181-189.
14. Wei, C.-P., Lee, Y.-H., & Hsu, C.-M. Empirical comparison of fast partitioning-based clustering algorithms for large data sets // Expert Systems with Applications, 24(4). – 2003. – P. 351-363.
15. Гаврилова Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Питер, 2001. – 384 с.

References

1. Diday, E. (1985), "*Metody analyza dannykh*" [*Optimisation en classification automatique*], Finance and Statistics, Moscow, 357 p.
2. Zhambi, M. (1988), "*Yerarxyčeskyj klaster-analyz y sootvetstvyja*" [*Hierarchical cluster analysis and matching*], Finance and Statistics, Moscow, 342 p.
3. Chernorutsky, I.G. (2005), "*Metody prynjatyja rešenyj*" [*Methods of decision making*], BHV-Petersburg, St. Petersburg, 416 p.
4. Barsegyan, A.A., Kupriyanov, M.S., Stepanenko, V.V. (2008), "*Texnolohyy analyza dannыx*" [*Data Analysis Technologies*], BHV-Petersburg, St. Petersburg, 384 p.
5. Zhdanov, A.A. (2008), "*Avtonomnyj yskusstvennyj yntellekt*" [*Autonomous artificial intelligence*], BINOM. Laboratory of Knowledge, Moscow, 359 p.
6. Bašmakov, A.Y., Bašmakov, Y.A. (2005), "*Yntellektual'nye ynformacyonnye texnolohyy*" [*Intellectual Information Technologies*], MGTU them. N.E. Bauman, Moscow, 304 p.
7. Burdaev, V.P. (2015), "*Slozhnostj dynamycheskykh system: monoghrafyja*" [Complexity of dynamic systems], LAP LAMBERT Academic Publishing, Saarbrucken, 140 p.
8. Burdaev, V.P. (2014), About one concept of constructing a temporal knowledge base, *Fundamental and Applied Studies in the Pacific and Atlantic Oceans Countries*: in Proc. of the 1st International Congress, Tokyo University Press, pp. 272-276.

9. Burdaev, V.P. (2009), "*Systemy navchannja z elementamy shtuchnogho intelektu: monoghrafija*" [*Systems of training with elements of artificial intelligence*], HNEU, Kharkiv, 392 p.

10. Burdaev, V.P. (2010), "*Modeli baz znanj: monoghrafija*" [*Models of knowledge bases*], HNEU, Kharkiv, 300 p.

11. Hae-Sang Park, Chi-Hyuck Jun. (2009), A simple and fast algorithm for K-medoids clustering, *Expert Systems with Applications*, 36, 336-334 p.

12. Khan, S.S., & Ahmad, A. (2004), Cluster center initialization algorithm for K-means clustering, *Pattern Recognition Letters*, 25, pp. 1293-1302.

13. Zhang, Q., & Couloigner, I. (2005), A new and efficient k-medoid algorithm for spatial clustering, *Lecture Notes in Computer Science*, 3482, pp. 181-189.

14. Wei, C.-P., Lee, Y.-H., & Hsu, C.-M. (2003). Empirical comparison of fast partitioning-based clustering algorithms for large data sets, *Expert Systems with Applications*, 24(4), 351–363p.

15. Gavrilova, T. and Khoroshevsky, V.F. (2001), "*Bazy znanyj yntellektual'nyx system*" [*Knowledge Base of Intelligent Systems*], Peter, St. Petersburg, 384 p.

Надійшла до редколегії 19.01.2018
Схвалена до друку 17.04.2018

Відомості про автора:

Бурдаєв Володимир Петрович
кандидат фізико-математичних наук
старший науковий співробітник
доцент кафедри
Харківського національного
економічного університету ім. Семена Кузнеця,
Харків, Україна
https://orcid.org/0000-0001-9848-9059

Information about author:

Volodymyr Burdayev
Candidate of Physics and Mathematics Sciences
Senior Research
Senior Lecturer of Department
Simon Kuznets Kharkiv National
University of Economics,
Kharkiv, Ukraine
https://orcid.org/0000-0001-9848-9059

СТРУКТУРА БАЗА ЗНАНИЙ ДЛЯ ВЫБОРА АЛГОРИТМА КЛАСТЕРИЗАЦИИ ДАННЫХ

В.П. Бурдаев

В статье обоснована актуальность задачи выбора алгоритма кластеризации объектов на основе знаний экспертов. Проблема кластеризации состоит в том, что для каждого конкретного типа данных, структуры расположения объектов в пространстве признаков надо или правильно подобрать известный алгоритм, или его адаптировать или разработать новый. Изучается математическая модель кластерного анализа. Адаптирован метод динамических сгущенний для размытой классификации разнородных данных. Излагается стратегия проведения кластеризации в системе "КАРКАС". Рассмотрены примеры правил и фреймов базы знаний для выбора алгоритма кластерного анализа.

Ключевые слова: *кластерный анализ, база знаний, экспертная система.*

STRUCTURE OF THE KNOWLEDGE BASE FOR CHOOSING THE DATA CLUSTERING ALGORITHM

V. Burdaev

The urgency of the classification problem without learning (clustering) for multidimensional objects of different nature is considered in the article. In clustering algorithms, the most important and least formalized is the definition of the concept of homogeneity or a measure of the closeness of objects, clusters and the quality of the partitioning of objects into groups (the objectivity of the groups obtained), which primarily determines the final result of the classification. All this indicates that the implementation of such algorithms in the form of application programs in batch mode is inefficient. Therefore, for an optimal heuristic solution of clustering tasks, the researcher must actively use the knowledge of experts on cluster analysis. The choice of this or that measure of the distance between clusters depends on the geometric figures that form objects in the space of characteristics. The application of the "nearest neighbor" distance has good clustering results when objects in the feature space form a chain. The distance of the "distant neighbor" is used when objects form ball clouds. In the case where objects form ellipsoids, it is recommended to use the distances between their centers of gravity. The problem of clustering is that for each particular type of data, the structure of the location of objects in the space of objects, you either need to choose the right algorithm correctly, either adapt it, or develop a new one. To solve this problem, experts' knowledge is widely used. The algorithm of the method of dynamic condensations for clustering heterogeneous data is proposed and adapted. Results are obtained for creating a knowledge base on cluster analysis. A knowledge base was built for the selection of algorithms: "K-intra-group means", "ISODATA", hierarchical and fuzzy clustering for different types of attributes. Examples of rules and frames that are used by the hierarchical functional system for making a decision on the choice of the clustering algorithm in the "KARKAS" system are given. To assess and compare the quality of the partitioning into clusters, different splitting quality functionals are used: "average intra-cluster scattering", "measure of the concentration of objects corresponding to the partitioning" and their combination. The results of clustering are presented: a table of distances between the centers of clusters, a table of variances to obtain an idea of the relative arrangement of images within the cluster.The knowledge base allows the expert to obtain additional information about the number, shape and compactness of clusters, the number of cluster centers and their coordinates, the distance between clusters and the dimension of "anomalous" clusters.

Keywords: *cluster analysis, knowledge base, expert system.*