УДК 004.9:519.23

ДВА ПІДХОДИ ДО РОЗВ'ЯЗАННЯ ЗАДАЧІ КЛАСТЕРИЗАЦІЇ У ШИРОКОМУ СЕНСІ З ПОЗИЦІЙ ІНДУКТИВНОГО МОДЕЛЮВАННЯ

В.В. Осипенко, кандидат технічних наук

Національний університет біоресурсів і природокористування України

Описано два оригінальних підходи до вирішення задачі кластеризації у постановці в широкому сенсі, базовані на застосуванні парадигми індуктивного підходу до кластер-аналізу. Методи можуть мати застосування у багатьох сферах прикладних системно-аналітичних досліджень, дотичних з проблемами структуризації, класифікації, кластеризації та моделюванням складних систем.

Кластеризація, критерій, цільова ознака, індуктивне моделювання.

У традиційних постановках задач кластерного аналізу "без учителя", як одного із важливих напрямків розпізнавання образів та інтелектуальних систем підтримки прийняття рішень (ІСППР), мають місце добре відомі некоректності, для подолання яких необхідно застосовувати численну апріорну інформацію разом із евристичними та інтуїтивними припущеннями стосовно наявної вибірки об'єктів, зокрема – таблиці даних експерименту. Відомо також, що при розв'язанні задач розпізнавання образів суттєву роль відіграє етап конструювання підпросторів інформативних ознак.

Це, природно, впливає на об'єктивність розв'язання задач моделювання. У цій роботі розглянуто два підходи до розв'язання задачі вибору інформативних ознак (ансамблю ознак) для цілей кластер-аналізу й отримання стійких кластеризацій вибірок даних спостережень, отриманих в активних чи пасивних експериментах, що не є принципово.

Міркування проводитимемо в межах постановки задачі так званої двоїстої (подвійної) кластеризації, або кластеризації у широкому сенсі, яка часто

зустрічається в економіці, екології, енергетиці, медичній діагностиці, біології і у багатьох інших прикладних напрямках.

Мета роботи – подання двох підходів до вирішення задачі кластеризації у широкому сенсі, як проблеми виявлення і формування гомогенних кластерних груп об'єктів із паралельним синтезом оптимальних ансамблів інформативних ознак, заснованих на методології індуктивного кластерного аналізу й із застосуванням цільової ознаки в конструкції критерію оцінки якості кластеризації, як регуляризуючого елементу.

Методи досліджень. Подані нижче підходи базуються на застосуванні принципів і методології індуктивного моделювання складних систем (ІМСС) до вирішення завдань кластеризації [1]. У цій роботі використані матеріали аналітичного огляду [2], методологія методу групового урахування аргументів (МГУА) [3], та посилання на загальновідомі [4] і сучасні алгоритми кластер-аналізу [5].

Результати досліджень. Оскільки, відмінність описаних нижче методів від традиційних [4] полягає у тому, що вони побудовані на застосуваннях принципів ІМСС (самоорганізації), тут будуть розглянуті також і критерії селекції [6] для вирішення задач кластеризації і відповідні схеми багаторядних алгоритмів.

ПОСТАНОВКА ЗАДАЧІ КЛАСТЕРИЗАЦІЇ В ШИРОКОМУ СЕНСІ

Нехай задано загальний масив вхідних даних в такому виді:

$$\tilde{X} = \left(x_{0j} \vdots x_{ij} \in X \right),\ j = \overline{1,m},\ i = \overline{1,n} \quad , \qquad (1)$$

де $x_{0j} \in (x_{01}, x_{02}, ..., x_{0m})$ – вектор цільових ознак, X – матриця вхідних ознак.

Тобто, кожен об'єкт (зображення) $\omega_j \in \Omega$ представлений описом виду

$$\omega_j = \left(x_{0j} \vdots x_{ij} \in X \right), i = \overline{1,n} \ .$$

Необхідно:

1) синтезувати підмножину $\left\{ x_\eta^* \right\} = X^* \subset X$, $\eta = 1, ..., n^*, n^* \le n$ із зазначених вище ознак, найкращу за заданим критерієм оптимальності та яка дозволила б:

2) класифікувати всі об'єкти з Ω на $k < m$, $k = 1, ..., K$ однорідних груп.

У термінах індуктивного підходу до завдань кластеризації підмножина $\{x_\eta^*\} = X^*$ зазвичай трактується ще як ансамбль інформативних ознак [2].

МЕТОДИ РОЗВ'ЯЗАННЯ ЗАДАЧІ

Критерії якості в індуктивному кластер-аналізі з цільовим ознакою.

Опишемо критерії оптимальності вирішення сформульованої вище в широкому сенсі задачі кластер-аналізу, які будуть застосовуватися в процедурах індуктивної кластеризації.

Відомо, що серед основних характеристик *k*-го кластера (для зручності і без втрати загальності далі будемо розглядати евклідів простір $\Re^N$) є його центр маси в просторі ознак *X*:

$$\bar{m}_k(X) = \left\{ \left(\frac{1}{r_k} \sum_{l=1}^{r_k} x_{l1} \right), \left(\frac{1}{r_k} \sum_{l=1}^{r_k} x_{l2} \right), \ldots, \left(\frac{1}{r_k} \sum_{l=1}^{r_k} x_{l\mathrm{n}} \right) \right\} =$$

$$= \left\{ \frac{1}{r_k} \sum_{l=1}^{r_k} x_{li},\ i = 1, ..., n \right\},\ x_i \in X \qquad , \qquad (2)$$

а середня внутрішньо-множинна відстань може бути подана як

$$\overline{d_k^2(\omega_s^k, \omega_t^k)} = \frac{1}{r_k (r_k - 1)} \sum_{s=1}^{r_k} \sum_{t=1}^{r_k} \sum_{i=1}^{n} \left(x_{is} - x_{it} \right)^2 \quad , \qquad (3)$$

де r_k – кількість вихідних об'єктів ω^k в *k*-му кластері, *n* – початкова кількість ознак простору *X*.

Використаємо цільову ознаку, елемент як регулюючий і обчислимо центр *k*-го кластера лише за значеннями x_0 об'єктів ω^k в *k*-му кластері. Це можна трактувати як проекцію центру *k*-го кластера з *n*-мірного евклідового простору *X* в одномірний евклідів простір $\Re^1$, тобто на вісь дійсних чисел.

Вирази (2) і (3) при цьому набувають більш простих виглядів:

$$\bar{m}_k(x_0) = \hat{m}_k = \frac{1}{r_k} \sum_{l=1}^{r_k} x_{0l} \qquad , \qquad (4)$$

$$\overline{d_k^2(\omega_s^k, \omega_t^k)_{x^0}} = \hat{d}_k^2 = \frac{1}{r_k(r_k-1)} \sum_{s=1}^{r_k} \sum_{t=1}^{r_k} \left(x_{0s} - x_{0t}\right)^2, \; s \neq t \; . \qquad (5)$$

Відомо, що застосування методології індуктивного моделювання складних систем [4] для одержання оптимальної кластеризації вимагає використання критеріїв, що мають властивості зовнішнього доповнення.

Підхід І. Цей підхід передбачає, щоб, використовуючи задану міру подібності, було встановлено таке розбиття (кластеризацію) вибіркової множини Ω незалежно в просторі цільових і вхідних ознак, щоб подвійні області R_k і $\dot{R}_k$ $k = 1, ..., K$, були "близькими" в сенсі заданого критерію якості розбиття з властивістю зовнішнього доповнення. При цьому, оскільки розглядається задача розпізнавання "без учителя", у початковій множині відсутні як об'єкти з індикаторами приналежності до якого-небудь кластеру, так і кількість самих кластерів.

Для цього позначимо опис зображення $\omega_j \in \Omega$ у просторі цільових ознак через D_1, а у просторі вхідних вимірювань – D_2.

Припустимо, що в просторі вхідних і вихідної ознак за деяким методом [5], [6] незалежно класифіковані об'єкти усієї множини $\omega_j \in \Omega$. Якщо за основну характеристику кластера прийняти його центр на осі цільової ознаки і задати, щоб сума квадратів відхилень центрів відповідних кластерів, встановлених на D_1 і D_2, була мінімальною, то критерій якості може бути представлений у такому вигляді:

$$\rho^2(\dot{m}) = \sum_{k=1}^{K} (\dot{m}_k(D_1) - \dot{m}_k(D_2))^2 \to \min \quad , \qquad (6)$$

де $\dot{m}_k(D_1)$ – центр k-го кластера в кластеризації за першим описом D_1 ; $\dot{m}_k(D_2)$ – центр k-го кластера в кластеризації за описом D_2, обчислений за значеннями цільової ознаки точок k-го кластера. Назвемо кластеризацію вздовж осі x_0 – *первинною*, а кластеризацію в просторі вхідних ознак – *вторинною*.

Для того, щоб критерій (6) змінювався в межах від 0 до 1, пронормуємо його відповідним чином і тоді він набуде вигляду:

$$\rho^2(\dot{m}) = \sum_{k=1}^{\hat{n}} (\dot{m}_k(D_1) - \dot{m}_k(D_2))^2 \,/\, \sum_{k=1}^{\hat{n}} (\dot{m}_k(D_1) + \dot{m}_k(D_2))^2 \to \min \qquad (7)$$

Вираз (7) називатимемо *критерієм найменших міжцентрових відхилень* (критерій НМВ) відносно осі x_0 (у просторі $\square^1$). Цей критерій можна також назвати критерієм несуперечності [1], [6] для задач кластеризації.

На рис. 1 показаний принцип роботи критерію $\rho^2(\dot{m})$ для прикладу з кількістю кластерів $K = 3$.

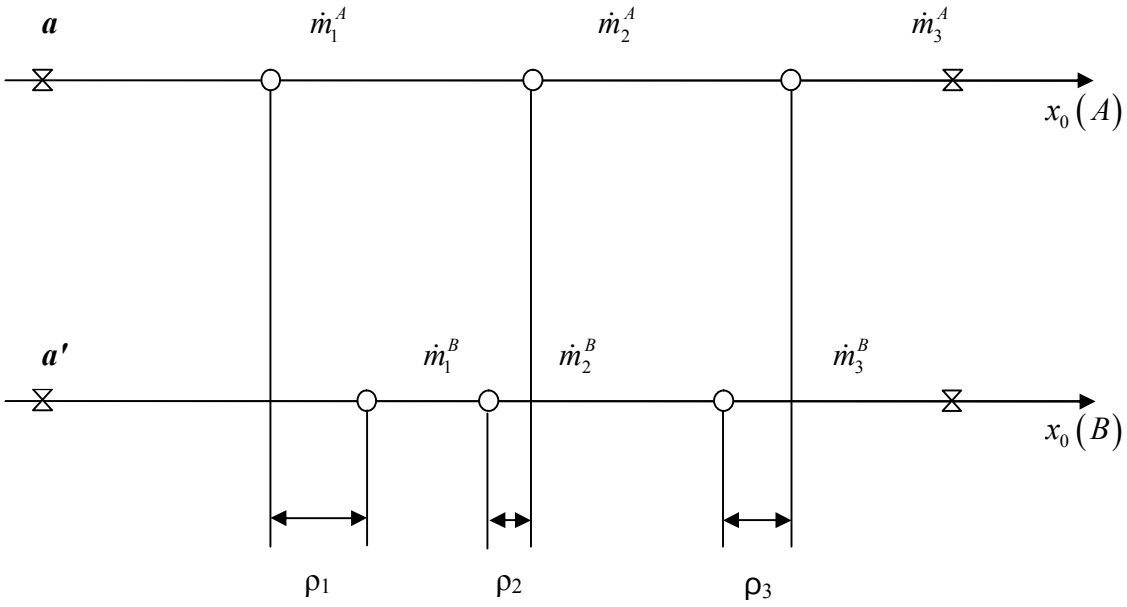

Рис. 1. До принципу роботи критерію найменших міжцентрових відхилень (несуперечності кластеризацій): $\rho_\Sigma = |\rho_1| + |\rho_2| + |\rho_3|)$

Підхід II. У цьому підході застосована класична схема розбиття вхідного масиву даних (1). Тобто, тут вимагається поділ вхідної множини об'єктів $\omega_k \in \Omega$, що підлягають кластеризації , не менше ніж на дві непересічні підмножини Ω^A і Ω^B, при цьому: $\Omega^A \cup \Omega^B = \Omega,\ \Omega^A \cap \Omega^B = \varnothing$.

Нехай на підмножинах Ω^A і Ω^B по одній з вибраних процедур кластер-аналізу, наприклад [4] чи [5], отримано кластеризації $s_t^A \in S^A$ і $s_t^B \in S^B$ з однаковими кількостями кластерів $k_t^A = k_t^B = K_t$ (t – номер кластеризації, що відповідає деякому підпростору ознак $X_t \subset X$, $k_t^{(\cdot)}$ – кількість кластерів в t-й кластеризації) в евклідовому підпросторі ознак $X_t \subset X$ і нехай для всіх

K_t кластерів із s_t^A і s_t^B обчислені їх центри $\dot{m}_k^A$ і $\dot{m}_k^B$, $k = 1,...,K_t$, по осі цільової ознаки x_0.

Тоді критерій оптимальності регуляризованої кластеризації можна записати в найпростішому і більш загальному вигляді як:

$$\rho^2(\dot{m}) = \sum_{k=1}^{\hat{K}} (\dot{m}_k^A - \dot{m}_k^B)^2 \to \min \quad . \tag{8}$$

Критерій (8) вимагає, щоб сума квадратів відхилень між центрами кластерів на осі цільового ознаки x_0, встановлених на подмножествах Ω^A і Ω^B, була мінімальною. Тому такий критерій у нашому випадку також називається критерієм найменших міжцентрових відхилень (НМВ), який, очевидно, теж є представником класу зовнішніх критеріїв несуперечності в методі групового урахування аргументів (МГУА). Аналогічно (7), для того, щоб значення критерію (6) змінювалися в межах деякого заданого інтервалу, наприклад [0, 1] , його можна записати у вигляді:

$$\rho^2(\dot{m}) = \sum_{k=1}^{K} (\dot{m}_k^A - \dot{m}_k^B)^2 \ / \ \sum_{k=1}^{K} (\dot{m}_k^A + \dot{m}_k^B)^2 \to \min \quad . \tag{9}$$

Це не є принциповим моментом, але така процедура часто застосовується в програмних реалізаціях алгоритмів МГУА. Принцип роботи цього критерію теж ілюструється за рис. 1 з відповідною інтерпретацією розбиття вхідного простору об'єктів.

Другий (допоміжний) критерій якості кластеризацій доцільно побудувати на класичному понятті мінімуму середніх внутрішньо множинних відстаней. У загальному випадку обчислення такого критерію за (3) є досить трудомісткою процедурою. Але в нашому підході ці відстані однозначно ідентифікуються по осі x_0, тобто в одномірному евклідовому просторі $\Re^1$, що істотно спрощує обчислення такого критерію, а саме:

$$\delta^2(\dot{d}^2) = \sum_{k=1}^{K} \dot{d}_k^2 \to \min \quad . \tag{10}$$

Критерій (10) не володіє властивістю зовнішнього доповнення, але характеризує якість кластеризацій, отриманих незалежно на підмножинах Ω^A і

Ω^B та може застосовуватися системно для більш чіткого відбору таких рішень, в яких уже отримані досить близькі значення критерію найменших міжцентрових відхилень – $\rho^2(\dot{m})$.

БАГАТОРЯДНІ АЛГОРИТМИ ІНДУКТИВНОЇ КЛАСТЕРИЗАЦІЇ

Підхід І. Наведена нижче версія алгоритму призначена для розв'язання задач подвійної кластеризації, коли цільова ознака і вхідні вимірювання пов'язані монотонними залежностями. Під монотонністю тут розуміємо той факт, що із нерівності $\|X_j\| \le \|X_{j+1}\|$ слідує нерівність $x_{0j} \le x_{0j+1},\ j = 1,\ldots,m-1$.

Алгоритм має таку схему обчислювальної багаторядної процедури.

1. Кластеризація множини Ω за описом D_1 (первинна кластеризація). При цьому задається деякий цілий параметр $K_r,\ 0 < K_r \le m$ – максимально можлива кількість кластерів. Первинну кластеризацію залежно від цілей і фізичних інтерпретацій даних можна виконати декількома способами, наприклад, за алгоритмом максимінної відстані [4], або за будь-яким іншим кластеризатором. В результаті на осі x_0 утворюється $K \le K_r$ груп зображень, для яких оцінюються середні значення $\overline{x_{0k}} = \dot{m}_k(D_1),\ k = 1,\ldots,K$. Частина зображень з кожної групи з мінімальними дисперсіями

$$\sigma_k^2 = \frac{1}{n_k - 1}\sum_{l=1}^{n_k}(x_{0lk} - \overline{x_{0k}})^2,\ k = 1,\ldots,K\ , \qquad (11)$$

відбираються в контрольну підмножину помічених об'єктів Δ_k (n_k – кількість зображень у *k*-му кластері).

2. Нормування всіх $\omega_i \in \Omega$ у просторі D_2 на інтервалі [0, 1]. Тут можна скористатися відомою формулою:

$$\tilde{x}_r = (x_r - x_{\min}^r) / (x_{\max}^r - x_{\min}^r)\ . \qquad (12)$$

На цьому кроці також виконується процедура генерування простору вторинних ознак $\{\hat{X}\}$ у заданому класі базисних наборів (функцій) F.

Примітка. У випадку дуже асиметричних даних доцільно замість евклідової використання канберрівської метрики:

$$d(\omega_{i,}\omega_j)=\sum_{s=1}^{N}\left[\left|x_{is}-x_{js}\right|/(x_{is}+x_{js})\right] \qquad (13)$$

У цьому випадку нормування початкових даних не потрібне, оскільки в результаті застосування (12) вони вже нормовані в необхідному інтервалі [0,1] за умови невід'ємності x_s.

3. Кластеризація (вторинна) множини $\Omega_f=\Omega\setminus\Omega_\Delta$ за ансамблями $\{\hat{\tilde{x}}_i\}$, $i=1,\ldots,N_1$ з використанням того ж алгоритму максиміну. У цьому алгоритмі формування $(k+1)$-го кластера починається за умови, якщо

$$\begin{aligned} &d\left(m_k,\ \omega_2\right)=[\sum_{i=1}^{n}(m_k-\hat{\tilde{x}}_i)^2]^{1/2}>\alpha r_t+\delta_r,\\ &\omega_l\in\Omega_f,\ t=1,\ldots,N_1,\ k=1,\ldots,K\ , \end{aligned} \qquad (14)$$

де $r_t=\sqrt{t}\,/\,K_{q,t}$, $-\frac{r_0}{2}\le\delta_r\le\frac{r_0}{2}$, $1<\alpha\le 2$, $K_{q,\mathrm{t}}$ – кількість встановлених кластерів у *q*-й кластеризації на *t*-му ряду селекції. Для всіх кластеризацій обчислюються значення критерію (7), які піддягають ранжуванню в порядку зростання. Згідно зі схемою багаторядних алгоритмів самоорганізації, зі всієї множини рішень на деякому ряді селекції відбирається тільки $F\ \le\ N$ кращих за критерієм (7), які і братимуть участь у подальшій селекції.

4. Кластеризація множини зображень Ω_f за ансамблями $\{\hat{\tilde{x}}_i^{(F)}\,\hat{\tilde{x}}_{\mathrm{j}}\}$, $j=1,\ldots,N_1$. Для цього використовується той же базовий алгоритм максимінної відстані і ті ж умови, що й на попередньому кроці. Також, як і вище, обчислюється критерій (7) і відбирається F рішень-претендентів.

Отже, якщо позначити *q*-й ансамбль з *F* вибраних на *t*-му ряді через $\{\hat{\tilde{X}}_t^q\}$, то на $(t+1)$-му ряді з нього формуватимуться ансамблі виду $\{\hat{\tilde{X}}_t^q,\hat{\tilde{x}}_i\}$, $i=1,\ldots,N_1$ з тими індексами, для яких $\hat{\tilde{x}}_i\in\hat{\tilde{X}}_t^q$.

Багаторядна процедура продовжується до тих пір, поки виконується умова:

$$\rho^2(\dot{m})_{\mathrm{t}}>\rho^2(\dot{m})_{\mathrm{t+1}}. \qquad (15)$$

Таким чином, результатом роботи описаної індуктивної процедури кластеризації у першому варіанті є:

1) оптимальна кластеризація $S^*(X^*)$ з отриманими гомогенними групами,

2) підпростір інформативних ознак $\{X^*\}$, за яким отримано оптимальну кластеризацію.

Підхід II. Цей алгоритм виконується за наступною процедурою.

Крок 1. Поділ (1) на дві частини А і В (Ω^A і Ω^B), згідно вимог методології індуктивного моделювання складних систем. Підготовлена загальна матриця даних $\tilde{X}$ буде мати такий умовний вигляд (припустимо, що *m* – парне):

$$\tilde{X} = \left[\left(x_{0j} \vdots X \right)^A \cdot \cdot \left(x_{0j} \vdots X \right)^B \right], \quad j = 1,...,m^A = m^B, \;\; m^A + m^B = m. \tag{16}$$

Крок 2. Налаштування однієї з процедур кластеризації (наприклад, класичного ієрархічного агломеративного алгоритму Ланса-Уільямса [4] або одного з сучасних алгоритмів типу [5] та ін.)

Крок 3. Кластеризація об'єктів $\omega_k \in \Omega$ за допомогою вибраного і уже налаштованого алгоритму незалежно на підмножинах Ω^A і Ω^B в просторі X за однією з класичних схем алгоритмів МГУА з індуктивним нарощуванням кількості ознак в їх ансамблях. Багаторядна індуктивна процедура кластеризації може бути такою.

1-й ряд селекції:

1.1) кластеризація об'єктів на підмножинах Ω^A і Ω^B за ансамблями $\{x_i\}$, $i = 1,...,n$;

1.2) проектування центрів отриманих кластерів на вісь x_0;

1.3) для кластеризацій, в яких виконується умова $k_t^A = k_t^B = K_t$ (t – поточний номер кластеризації, $k_t^{(\cdot)}$ – кількість кластерів в *t*–й кластеризації), обчислюються значення критерію оптимальності $\rho^2(\dot{m})$.

2-й ряд селекції:

2.1) кластеризація об'єктів на підмножинах Ω^A і Ω^B за ансамблями $\{x_i, x_j\},\ i, j = 1,...,n,\ i \neq j$;

2.2) виконуються п.п. (1.2) – (1.3) і за критерієм (9) відбираються F $(F \leq n)$ кращих кластеризацій S_f та відповідних ансамблів ознак X_f, $f = 1,...,F$.

3-й і наступні ряди селекції:

3.1) кластеризація об'єктів на підмножинах Ω^A і Ω^B за ансамблями $\left\{X_f, x_l\right\},\ f = 1,...,F,\ l = 1,...,n$ за умови, що ознака з індексом l не присутня в уже створених ансамблях X_f.

3.2) виконується п. (2.2).

Правило зупинки: індуктивна процедура зупиняється за умови:

$$\rho^2(\dot{m})_{\text{s}} \leq \rho^2(\dot{m})_{\text{s+1}}\ , \qquad (17)$$

де s – ряд селекції в термінах МГУА. При цьому фіксується значення $k^{*(A)} = k^{*(B)} = K^*,\quad K^* \leq m/2$ і підпростір інформативних ознак $\left\{x_l^*\right\} = X^*,\ l = 1,...,n^*,\ n^* \leq n$.

Отже, й у другому варіанті вирішення задачі кластеризації в широкому сенсі маємо синтезовану підмножину $\left\{x_\eta^*\right\} = X^* \subset X, \eta = 1, \ldots, n^*, n^* \leq n$ із усіх заданих з експерименту ознак, що є найкращою за заданим критерієм оптимальності і яка дозволяє класифікувати всі об'єкти з Ω на $k < m$, $k = 1, ..., K$ однорідних груп.

Окрім описаних вище критеріїв, у багаторядному алгоритмі для вибору ефективного ансамблю $\{X^*\}$ та підвищення надійності розпізнавання застосовується також такий тип критеріїв, як *критерій мінімуму помилки розпізнавання* (критерій МПР).

Сконструюємо такий критерій стосовно наведених алгоритмів. Оскільки ці алгоритми вирішують задачу кластеризації в однакових постановках, то й критерій мінімуму помилки розпізнавання для них будуть однаковим. Розпишемо такий критерій для, наприклад, підходу І. Для цього припустимо,

що рішення про належність контрольного зображення $\omega_l^* \in \Omega$ до якого-небудь кластеру береться на основі деякої міри близькості, який сформулюємо так: об'єкт ω_l^* належить до *k*-го кластера, якщо

$$d(\omega_l^*, m_k) < d(\omega_l^*, m_s), \;\; k,s = 1,\ldots,K, \;\; k \neq s \qquad (18)$$

для деякого ансамблю ознак $\{X_p^*\}$ і відповідній кластеризації S_p^*, де $(X_p^*, S_p^*), \;\; p = 1,\ldots$ – рішення з оптимальної області. На осі цільової ознаки вираз (18) набуде вигляду:

$$d(x_{0l}^*, \dot{m}_k) < d(x_{0l}^*, \dot{m}_s), \; k,s = 1,\ldots,K, \;\; k \neq s \; . \qquad (19)$$

Тоді критерій мінімуму помилки розпізнавання $\Delta^2(\Omega_\Delta)$ запишеться як:

$$\Delta^2(\Omega_\Delta) = \sum_{\Omega_\Delta} \delta_l^2 \to \min, \qquad (20)$$

де $$\delta_l = \begin{cases} 1, \; d(x_{0l}^*, \dot{m}_{k(l)}) > d(x_{0l}^*, \dot{m}_{s(l)}) \\ 0, \; d(x_{0l}^*, \dot{m}_{k(l)}) < d(x_{0l}^*, \dot{m}_{s(l)}) \end{cases}, \qquad (21)$$

де: $k \neq s, k, \; s = 1,\ldots,K$, $\omega_l^* \in R_k$, або:

$$\delta_l = [x_{0l}^* - \dot{m}_{k(l)}(D_2)], \;\; \omega_l^* \in \Omega_\Delta, \qquad (22)$$

де x_{0l}^* – значення цільової ознаки для $\omega_l^* \in \Omega_\Delta$; $\dot{m}_{k(l)}(D_2)$ – оцінка центру *k*-го кластера на осі x_0, до якого зарахований об'єкт ω_l^* по ансамблю $\{X_p^*\}$.

Рівність $d(x_{0l}^*, \dot{m}_{k(l)}) = d(x_{0l}^*, \dot{m}_{s(l)}), \; k \neq s$ при $\omega_l^* \in R_k$ відповідає ситуації "відмова від розпізнавання". У даному алгоритмі у такому випадку питання про остаточний вибір з'ясовується на користь ансамблю з меншою складністю і відповідною кластеризації, а на даному ряду – на користь R_k.

Крім того, на основі отриманого рішення можна розпізнати значення вихідної величини x_0^* *нового* (що не бере участь у самонавчанні) зображення:

$$\min_K d[(m_k, \omega^*) / S^*(X^*)] \to x_0^*. \qquad (23)$$

Такі підходи й відповідні багаторядні алгоритми кластеризації доцільно застосовувати для знаходження ефективної структури регресійної моделі в заданому класі *F* функцій, як показано нижче на модельному прикладі. Це

дозволяє істотно скоротити витрати часу комп'ютерних обчислень, оскільки при цьому немає потреби у розв'язанні складних систем рівнянь й обробці інших трудомістких операцій, властивих регресійному аналізу. З другого боку, деякі труднощі в застосуванні описаних вище алгоритмів для вибору ефективної структури моделі пов'язані з вибором адекватних мір подібності. Ці питання повинні розглядатись у кожному практичному випадку окремо.

Роботу індуктивного алгоритму кластеризації у широкому сенсі проілюструємо на прикладі із застосуванням описаного алгоритму I.

Приклад 7.1. Вибіркова матриця даних наведена в табл. 1. Припустимо, що шуканий ансамбль ознак – $\{X\} = \{x_1, x_2^2\}$ і $x_0 = x_1 + x_2^2$, причому кількість кластерів у просторі вхідних вимірювань невідоме.

Згідно постановки задачі кластер-аналізу у широкому сенсі, необхідно визначити:

- оптимальну кількість кластерів – K^*;
- підпростір інформативних ознак – $\{X^*\}$;
- відповідний $\{X^*\}$ кластеризації об'єктів – $S^*(X^*)$.

Простір вторинних ознак згенеруємо таким: $\{x_i, x_i^k\}$, $k = 1, 2, 3$, $i = 1, 2$. Значення ознак нормуємо в інтервалі [0, 1] за формулою (12). Значення параметра K_r будемо задавати із множини $K_r = \{2, 4, \ldots, 12\}$, а параметр δ_r візьмемо рівним 0,05, що є допустимим для практичних задач кластеризації.

Як неважко переконатися, у результаті застосування процедури кластеризації у реалізації першого підходу із застосування базового алгоритму максимінної відстані знаходяться три рішення з однаковими значеннями критерію (7), а саме $\rho^2(\dot{m}) = 0$. Це ансамблі: $\{x_1, x_2\}$, $\{x_1, x_4\}$, $\{x_2, x_3\}$.

Але легко перевірити, що використання критерію мінімуму помилки розпізнавання у вигляді (20), (22) дозволяє все-таки із трьох вибрати один оптимальний ансамбль і відповідну кластеризацію $S^*(X^*)$ за ансамблем $\{X^*\} = \{x_1, x_4\} \equiv \{x_1, x_2^2\}$.

Вихідні дані для прикладу [7]

№ об'єкту ω_i	x_0	x_1	x_2	№ кластеру, R_k	Розпізнавання (прогноз) x_0^*
1	4	3	1	1	
2	8	7	1	3	
3	13	9	2	5	
4	6	2	2	2	
5	10	1	3	4	
6	14	5	3	6	
7	20	4	4	7	
8	34	9	5	9	
9	28	3	5	8	
10	42	6	6	10	
11	54	5	7	11	
12	66	2	8	12	
ω_1^*	*36*	*6*	*5.5*	*9*	34 / *36*
ω_2^*	*72*	*8*	*8*	*12*	66 / *72*

Тепер уже легко розпізнати "нові" об'єкти ω_1^* і ω_2^* і визначити "невідомі" для них значення x_0, зарахувавши їх до кластерів, для яких така інформація уже відома. Отже, для ω_1^* значення $x_0^* = 34$ (реальне значення – 36) і для ω_2^* – $x_0^* = 66$ (реальне значення – 72).

Висновки

Задача кластеризації в широкому сенсі досить часто застосовується для попереднього аналізу "сирих" даних з метою виявлення компактних груп досліджуваних об'єктів і, що важливо, конструювання оптимальних в сенсі сконструйованих критеріїв факторних підпросторів. Такий підхід, зокрема, часто застосовується як попередній етап перед застосуванням більш "витончених" інструментів, таких як, наприклад, параметрична ідентифікація. Це дозволяє істотно скоротити об'єми обчислень на етапах структурної ідентифікації, оскільки попередньо уже синтезовано ансамбль $\{X^*\}$.

У роботі описані два підходи до вирішення задачі кластеризації саме у поданій постановці. Обидва підходи базовані на застосуванні методології індуктивного підходу до кластер-аналізу. Методи можуть мати застосування у багатьох сферах прикладних системно-аналітичних досліджень, дотичних з

проблемами структуризації, класифікації, кластеризації та моделюванням складних систем.

Список літератури

1. Ивахненко А.Г. Объективная кластеризация на основе теории самоорганизации моделей / А.Г. Ивахненко // Автоматика. – 1987. – №5. – С. 6 – 15.
2. Сарычева Л.В. Объективный кластерный анализ данных на основе МГУА / Л.В. Сарычева // Проблемы управления и информатики. – 2008. – № 2. – С. 86 –104.
3. Ivakhnenko A.G. Inductive Learning Algorithms for Complex Systems Modeling / R. Madala, A.G. Ivakhnenko. – CRC Press, Boca Raton, 1994. – 350 p.
4. Мандель И. Д. Кластерный анализ / И. Д. Мандель. – М.: Финансы и статистика. – 1988. – 176 с.
5. Литвиненко В.И. Кластерный анализ данных на основе модифицированной иммунной сети / В.И. Литвиненко // Управляющие системы и машины. – УСиМ. – 2009. – №1. – С. 54 – 61, 85.
6. Степашко В.С. Елементи теорії індуктивного моделювання / Стан та перспективи розвитку інформатики в Україні: монографія / Колектив авторів. – Київ: Наукова думка, 2010. – 1008 с. / – С. 471 – 486.
7. Осипенко В.В. Решение задачи двойной кластеризации на основе самоорганизации / В.В. Осипенко //Автоматика. –1988. – №3. – С. 74 – 79. [Osypenko V. Solution of a Double Clusterization Problem with the Use of Self-Organization. In: SAC, USA, No. 3, vol.21, –1988. – Pp.77 – 82.

Описанные два оригинальных подходы к решению задачи кластеризации в постановке в широком смысле, основанные на применении парадигмы индуктивного подхода к кластер-анализа. Методы могут иметь применение во многих сферах прикладных системно-аналитических исследований, соприкасающихся с проблемами структуризации, классификации, кластеризации и моделированием сложных систем.

Кластеризация, критерий, целевой признак, индуктивное моделирование.

Described two original approaches to the problem of clustering in the formulation in a broad sense, based on the application of the paradigm of inductive approach to cluster analysis. Methods can be used in many areas of applied system-analytical studies in contact with the problems of structuring, classification, clustering, and modeling of complex systems.

Clustering, criterion, target indication, inductive simulations.