УДК 681.518:004.93.1’

# КЛАСТЕРИЗАЦІЯ ВХІДНИХ ДАНИХ

А. С. Довбиш, д-р техн. наук; С. П. Пашко, студент,
Сумський державний університет
kichrum@dl.sumdu.edu.ua

Використання кластер-аналізу в рамках інформаційно-екстремальної інтелектуальної технології (IEI-технології) [1] дозволяє автоматизувати процес формування апріорно класифікованої навчальної багатовимірної матриці. Розглянемо кластеризацію вхідних даних на прикладі формування навчальної матриці для системи контролю рівня знань студентів за навчальною дисципліною «Інтелектуальні системи», що викладається студентам спеціальності «Інформатика» в Сумському державному університеті. За результатами тестування рівня знань студентів було одержано 488 реалізацій чотирьох класів (клас $X_5^o$ – «відмінно», клас $X_4^o$ – «добре», клас $X_3^o$ – «задовільно», клас $X_2^o$ – «незадовільно»). Загальна кількість тестів, яка визначала потужність словника ознак розпізнавання, дорівнювала $N = 141$, тобто за ознаку розпізнавання брали результат відповіді студента на відповідний тест, що відображався за обраною оцінною функцією на стобальну шкалу. Введемо обмеження, які спрощують задачу кластеризації вхідних даних: потужність алфавіту класів є обмеженою і дорівнює $Card\{X_m^o\} = 4$; алфавіт класів розпізнавання є впорядкованим, тобто двійковий еталонний вектор класу $X_2^o$ є найближчим до вершини нульового вектора-реалізації (значення всіх ознак знаходяться поза своїми контрольними допусками), й еталонний вектор класу $X_5^o$ є найближчим до вершини одиничного вектора-реалізації (значення всіх ознак знаходяться у своїх контрольних допусках, оскільки всі відповіді на тести були правильними). Інформаційно-екстремальний алгоритм навчання СППР із кластеризацією вхідних даних полягає в перетворенні неструктурованої вхідної навчальної матриці $\| y_i^{'(j)} \mid i = \overline{1, N};\ j = \overline{1, n} \|$ в апріорно нечітку класифіковану багатовимірну навчальну матрицю $\| y_{m,i}^{(j)} \mid m = \overline{1, M};\ i = \overline{1, N};\ j = \overline{1, n} \|$ та відображенні її в дискретний (субпарацептуальний) простір ознак розпізнавання, де шляхом допустимих цілеспрямованих перетворень вхідний математичний опис адаптується з метою максимізації повної ймовірності правильного прийняття рішень.

Розглянемо схему ієрархічного алгоритму кластер-аналізу вхідних даних для формування нечіткої класифікованої багатовимірної навчальної матриці. Спочатку побудуємо контейнери для двох класів розпізнавання, що знаходяться на верхньому рівні ієрархічної структури.

Крок 1. Формуються двійковий одиничний вектор $x_5^{(1)}$ і аналогічно – нульовий вектор $x_2^{(0)}$, які за структурою відповідають векторам-реалізаціям класів розпізнавання.

Крок 2. Обнуляється лічильник кроків зміни радіуса контейнера відповідного класу, що відновлюється в радіальному просторі ознак розпізнавання: $r := 0$.

Крок 3. Ініціалізація лічильника кроків прирощення радіуса: $r := r + 1$.

Крок 4. Біля вершини вектора $x_5^{(1)}$ будується таксон $T_5^{(1)}$ радіуса $r$.

Крок 5. Якщо для будь-якого вектора $x^{(j)}$ спостерігається $x^{(j)} \in T_5^{(1)}$, то виконується крок 6. Інакше – крок 3.

Крок 6. За дистанційною мірою $d[x_5^{(1)} \oplus x^{(j)}]$ у таксоні визначається найближчий до одиничного вектор $x_{5,\min}^{(j)}$, вершину якого беруть за центр нового таксона $T_5^{'}$, і виконується крок 2.

Аналогічно знаходиться вектор $x_{2,\min}^{(j)}$, найближчий до нульового, вершину якого беруть за центр нового таксона $T_2^{'}$. Далі для кожного з таксонів $T_2^{'}$ і $T_5^{'}$ запускається агломеративний алгоритм пошуку відповідних центрів ваги. При цьому відбувається ініціалізація лічильника кроків прирощення радіусів таксонів, яка припиняється за умови $r \le d[x_{2,\min} \oplus x_{5,\min}]/2$. Використання такої умови дозволяє побудувати на верхньому ієрархічному рівні таксони класів $X_2^{'}$ і $X_5^{'}$, які містять усі вектори-реалізації із заданого розподілу.

При переході на нижній рівень ієрархічної структури для побудови таксона класу $X_2^o$ було використано як початкову реалізацію $x_{2\min}$, а для побудови таксона класу $X_5^o$ – реалізацію $x_{5\min}$, оскільки ці реалізації з найбільшою ймовірністю належать до відповідних класів. На радіуси таксонів класів $X_2^o$ і $X_5^o$ відповідно накладалися такі обмеження:

$$r_2 \le \frac{d[x_2 \oplus x_2^{(0)}]}{2}; \qquad r_5 \le \frac{d[x_5 \oplus x_5^{(1)}]}{2}. \qquad (1)$$

Для агломерації реалізацій класів $X_3^o$ і $X_4^o$ як початкові обиралися вектори, вершини яких належали відповідно класам $X_2^{'}$ і $X_5^{'}$, але не належали контейнерам класів $X_2^o$ і $X_5^o$. Радіуси таксонів для класів $X_3^o$ і $X_4^o$ відповідно дорівнювали максимальним радіусам (1) таксонів класів $X_2^o$ і $X_5^o$.

Таким чином, за умови заданої потужності структурованого алфавіту класів розпізнавання вдалося за дистанційними критеріями побудувати нечітке розбиття простору ознак на класи розпізнавання, що дозволило сформувати вхідну нечітку класифіковану навчальну матрицю. Реалізація алгоритму навчання у рамках IEI-технології з використанням сформованої навчальної матриці дозволила трансформувати апріорно нечітке розбиття простору ознак на класи розпізнавання в чітке розбиття, тобто побудувати безпомилкові за навчальною матрицею вирішальні правила.

1. Довбиш А. С. Основи проектування інтелектуальних систем: навчальний посібник / А. С. Довбиш.– Суми: Видавництво СумДУ, 2009.– 171 с.