Кластерный анализ

Пример 2. На основании данных, приведенных ниже, проведите классификацию магазинов по трем признакам: X_1 – площадь торгового зала, м^2 , X_2 – товарооборот на одного продавца, ден. ед., X_3– уровень рентабельности, %.

Номер магазина	X_1	X_2	X_3	Номер магазина	X_1	X_2	X_3
1	100	160	25	6	85	200	35
2	130	200	30	7	60	170	28
3	80	180	20	8	110	150	18
4	40	100	22	9	55	110	15
5	150	90	15	10	110	100	12

Решение. 1. Рассчитаем расстояния между объектами по евклидовой метрике

$$d_{ij} = \sqrt{\sum_{i=1}^{m} \left(z_{ik} - z_{jk} \right)^2},$$

где z_{ik}, z_{jk} – стандартизированные значения исходных переменных соответственно у i-го и j-го объектов; m – число признаков.

$$z_{ij} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$

$$Z = \begin{pmatrix} 0{,}243 & 0{,}345 & 0{,}426 \\ 1{,}156 & 1{,}338 & 1{,}136 \\ -0{,}365 & 0{,}838 & -0{,}284 \\ -1{,}582 & -1{,}134 & 0{,}000 \\ 1{,}764 & -1{,}381 & -0{,}994 \\ -0{,}273 & 1{,}332 & 1{,}847 \\ -0{,}973 & 0{,}592 & 0{,}852 \\ 0{,}547 & 0{,}099 & -0{,}568 \\ -1{,}125 & -0{,}888 & -0{,}994 \\ 0{,}547 & -1{,}134 & -1{,}420 \end{pmatrix}$$

2. На основе матрицы Z рассчитаем квадратную симметричную матрицу расстояний между объектами (D).

$$D = \begin{bmatrix} 0 & 1{,}524 & 1{,}052 & 2{,}609 & 2{,}324 & 1{,}805 & 1{,}312 & 1{,}069 & 2{,}325 & 2{,}385 \\ & 0 & 2{,}140 & 3{,}860 & 3{,}507 & 1{,}596 & 2{,}274 & 2{,}193 & 3{,}833 & 3{,}608 \\ & & 0 & 2{,}335 & 3{,}156 & 2{,}189 & 1{,}312 & 1{,}208 & 2{,}015 & 2{,}452 \\ & & & 0 & 3{,}499 & 3{,}348 & 2{,}019 & 2{,}525 & 1{,}121 & 2{,}559 \\ & & & & 0 & 4{,}425 & 3{,}846 & 1{,}963 & 2{,}931 & 1{,}313 \\ & & & & & 0 & 1{,}424 & 2{,}833 & 3{,}705 & 4{,}175 \\ & & & & & & 0 & 2{,}138 & 2{,}371 & 3{,}233 \\ & & & & & & & 0 & 1{,}988 & 1{,}499 \\ & & & & & & & & 0 & 1{,}743 \\ & & & & & & & & & 0 \end{bmatrix}$$

3. Зададим радиус сферы $R = 2$. В этом случае в сферу попадают объекты, расстояние которых до первого объекта меньше 2.

$$d_{12} = 1{,}524 \qquad d_{13} = 1{,}052 \qquad d_{16} = 1{,}805 \qquad d_{17} = 1{,}312 \qquad d_{18} = 1{,}069$$

Для шести точек (объекты 1, 2, 3, 6, 7, 8) определяем координаты центра тяжести: $\bar{x}_* = (0{,}056;\ 0{,}757;\ 0{,}568)$

4. На следующем шаге алгоритма помещаем центр сферы в точку $\bar{x}_*$ и определяем расстояние каждого объекта до нового центра:

$$d_{1*} = 0{,}474 \qquad d_{2*} = 1{,}367 \qquad d_{3*} = 0{,}954 \qquad d_{4*} = 2{,}565 \qquad d_{5*} = 3{,}151$$

$$d_{6*} = 1{,}440 \qquad d_{7*} = 1{,}080 \qquad d_{8*} = 1{,}401 \qquad d_{9*} = 2{,}557 \qquad d_{10*} = 2{,}787$$

Следовательно, в сферу попали объекты 1, 2, 3, 6, 7, 8, расстояния которых до центра меньше радиуса сферы. Поскольку в этом случае центр сферы не изменит своих координат, выделение первого кластера закончено, в его состав вошли шесть объектов (1,2,3,6,7,8).

5. Чтобы начать формирование второго кластера, нужно поместить центр сферы в одну из точек, не вошедших в первый кластер (объекты 4,5,9,10).

Судя по матрице расстояний , целесообразно в качестве центра сферы выбрать объекты 9 или 10. Если взять объект 9 в качестве центра сферы, то в сферу попадают четыре точки (объекты 4,8,9,10). Рассчитаем для них координаты нового центра тяжести .

6. Определим расстояние каждого из десяти объектов до точки :

$\bar{x}_*$

$$d_{1*} = 1{,}738 \qquad d_{2*} = 2{,}646 \qquad d_{3*} = 1{,}668 \qquad d_{4*} = 1{,}443 \qquad d_{5*} = 3{,}151$$

$$d_{6*} = 1{,}888 \qquad d_{7*} = 2{,}172 \qquad d_{8*} = 1{,}296 \qquad d_{9*} = 1{,}888 \qquad d_{10*} = 1{,}222$$

В сферу попадают объекты, которые имеют расстояние до центра меньше двух (объекты 1,3,4,8,9,10).

На основании матрицы Z по евклидовой метрике определяем новые координаты центра для этих точек $\bar{x}_* = (-0{,}289;\ -0{,}312;\ -0{,}473)$.

Для нового центра повторяем пункт 6 данного алгоритма:

$$d_{1*} = 1{,}234 \qquad d_{2*} = 2{,}720 \qquad d_{3*} = 1{,}379 \qquad d_{4*} = 1{,}603 \qquad d_{5*} = 2{,}373$$

$$d_{6*} = 2{,}843 \qquad d_{7*} = 1{,}744 \qquad d_{8*} = 0{,}936 \qquad d_{9*} = 1{,}141 \qquad d_{10*} = 1{,}507$$

После выполнения этого шага видно, что в сферу с радиусом R=2 попадают объекты 1,3,4,7,8,9,10, т.е. состав второго кластера опять изменился. Следовательно, повторяются процедуры пункта 6 и пункта 7:

$$\bar{x}_* = (-0{,}387;\ -0{,}183;\ -0{,}284)$$

$$d_{1*} = 1{,}086 \qquad d_{2*} = 2{,}590 \qquad d_{3*} = 1{,}021 \qquad d_{4*} = 1{,}553 \qquad d_{5*} = 2{,}562$$

$$d_{6*} = 2{,}617 \qquad d_{7*} = 1{,}495 \qquad d_{8*} = 1{,}016 \qquad d_{9*} = 1{,}243 \qquad d_{10*} = 1{,}751$$

Как видно из полученных расстояний каждого из десяти объектов до центра второго кластера, состав кластера не изменился. На этом выделение второго кластера завершается. В его состав вошли семь объектов 1,3,4,7,8,9,10.

Результаты выделения первых двух кластеров представлены на рисунке 1

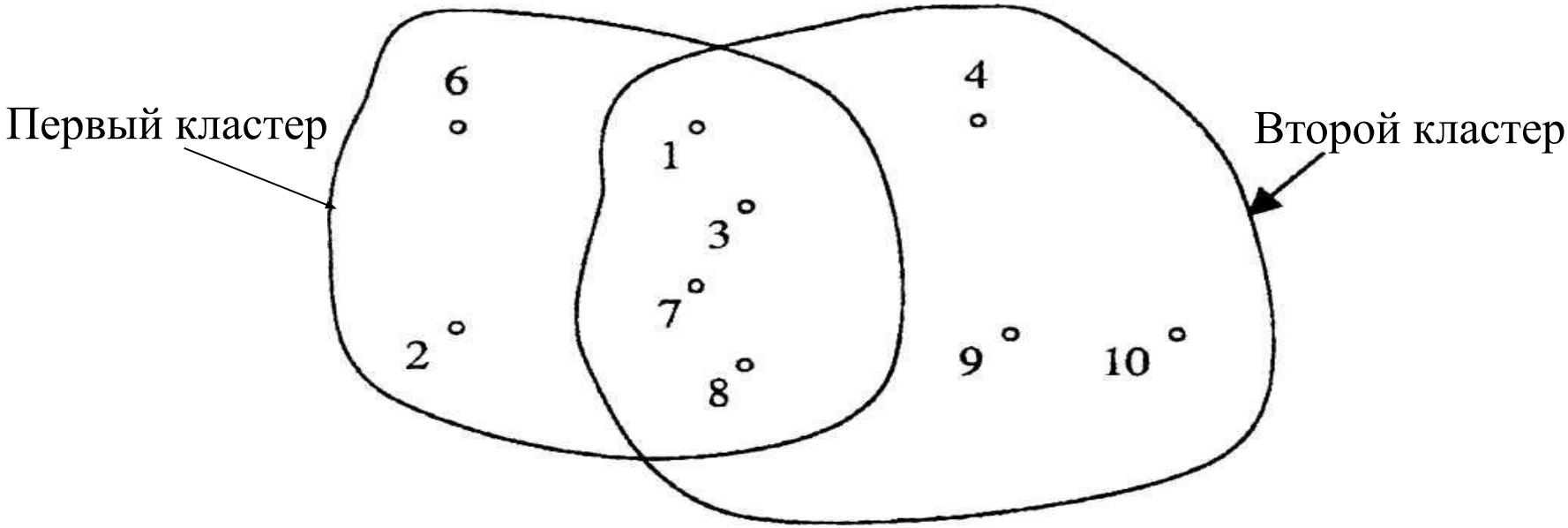

Рисунок 1 – Два выделенных пересекающихся кластера