Интерактивный самоорганизующийся метод анализа данных – ИСОМАД

- Выбор исходных управляющих параметров
- Условия соединения кластеров
- Условия расщепления кластеров
- Алгоритм распознавания (БС)

Алгоритм ИСОМАД является интерактивным многопараметрическим алгоритмом. В качестве эталонов кластеров в нем служат выборочные средние, определяемые на каждой итерации.

Схема алгоритма представляет широкие возможности применения "эвристических" процедур при модификации.

Шаг 1. В качестве начальных исходных данных (управляющих параметров) следует задать:

$S = S_1, S_2, \ldots, S_m$ - множество допустимых объектов;

$Z_1, Z_2, \ldots, Z_{l_1}$ - исходное число центров, полученное из S;

l - необходимое число кластеров, необязательно равное l_1;

$_m$ - параметр, с которым сравнивается количество объектов, вошедших в кластер;

$_s$ - параметр, характеризующий среднеквадратичное отклонение;

$_c$ – параметр, характеризующий компактность кластеров;

L_z – максимальное количество пар центров кластеров, которое можно объединить;

ITR – число итераций.

Шаг 2. Заданные объекты S_i множества S, распределяются по кластерам, соответствующим выбранным исходным центрам Z_j,

$(j=1, 2, \ldots, l_1)$ по правилу

$S_i \; K_t$, если $(S_i, Z_t) \quad (S_i, Z_j), j=1, 2, \ldots, l_1, i=1, 2, \ldots, m.$

Через K_j, $j=1, 2, \ldots, l_1$ обозначены кластеры, центрами которых являются соответственно объекты

Z_j, $(j=1, 2, \ldots, l_1)$.

Шаг 3. Ликвидируются кластеры, в состав которых входит менее $_m$ объектов, т.е. если для некоторого j выполняется условие $m_j \; _m$, то кластер K_j с числом объектов m_j исключается из рассмотрения как кластер и значение l_1 уменьшается на 1.

Шаг 4. Если какой-то кластер расформирован $(l_1=l_1-1)$, то объекты, входящие в такой кластер распределяются по остальным кластерам (центрам) по ближайшему соседу, аналогично шага 2 по правилу

$S_p \; K_t$, если $(S_p, Z_t) \quad (S_p, Z_j)$, где

S_p– объекты расформируемого объекта, $j=1, 2, \ldots, l_1-1, t \; j.$

Здесь в качестве метрики близости можно также использовать расстояние от объекта S_p до кластера K_j, (S_p, K_j).

Шаг 5. Каждый центр Z_j кластера K_j, $j=1, 2, \ldots, l_1$ локализуется и корректируется путем приравнивания его выборочному среднему, найденному по соответствующему кластеру K_j, т.е.

$$Z_j = \frac{1}{m_j} \sum_{S_i \in K_j} S_i, \; j=1,2,\ldots,l_1, \tag{2.33}$$

где m_j - число объектов, вошедших в кластер K_j.

Шаг 6. Вычисляется среднее расстояние d_j между объектами, входящими в подмножество K_j, и соответствующим центром кластера по формуле

$$d_j = \frac{1}{m_j} \sum_{S \in K_j} \rho(S, Z_j), \; j=1,2,\ldots,l_1. \tag{2.34}$$

Шаг 7. Вычисляется обобщенное среднее расстояние между объектами, находящимися в отдельных кластерах, и соответствующими центрами кластеров по формуле

$$d=\frac{1}{m}\sum_{j=1}^{l_1} N_j d_j, N_j=\left|k_j\right|. \qquad (2.35)$$

Шаг 8.

а) Если текущий цикл итерации - последний, то задается $_c=0$. Переход к шагу 12;

б) Если условие $l_1 \le \frac{l}{2}$ выполняется, то переход к шагу 9;

в) Если текущий цикл итерации имеет четный порядковый номер или выполняется $m \quad 2l$, то переход к шагу 12. В противном случае, процесс итерации повторяется.

Шаг 9. Для каждого кластера с помощью соотношения вычисляется вектор среднеквадратического отклонения

$_j=(\ _{1j},\ _{2j},\ldots,\ _{nj}\ )$:

$$\sigma_{ij}=\sqrt{\frac{1}{N_j}\sum_{S_k \in K_j} ¿¿¿} \qquad (2.36)$$

где n – число признаков, характеризующих объект; S_{ik} - есть i– й признак k–го объекта в кластере K_j, Z_{ij} есть i–я компонента вектора, представляющего центр кластера Z_j, N_j – количество объектов, включенных в кластер K_j. Каждая компонента вектора среднеквадратичного отклонения $_j$ характеризует среднеквадратичное отклонение объекта, входящего в K_j, по одной из главных осей координат (по одному признаку).

Шаг 10. В каждом векторе среднеквадратического отклонения $_j$ $j=1, 2,..., l_1$, отыскивается максимальная компонента $_{j\max}$.

Шаг 11. Если для любого $_{j\max}$ $j=1, 2,..., l_1$ выполняются условия $_{j\max} \quad _s$ и

а) $d_j \quad d$ и $m_j \quad 2\ (\ _m+1)$ или

б) $l_1 \le \frac{l}{2}$, (2.37)

то кластер с центром Z_j расщепляется на два новых кластера с центрами Z_j^- и $Z_j^{+¿¿}$ соответственно, кластер с центром Z_j ликвидируется, а значение l_1 увеличивается на единицу. Для определения центра кластера $Z_j^{+¿¿}$ к компоненте вектора Z_j, соответствующей максимальной компоненте вектора $_j$, прибавляется заданная величина $_j$. Центр кластера Z_j^- определяется вычитанием этой же величины $_j$ из той же самой компоненты вектора Z_j. В качестве величины $_j$ можно выбрать некоторую долю значения максимальной среднеквадратичной компоненты $_{j\max}$, т.е. положить $_j=k\ _{j\max}$, $0 \quad k \quad 1$. При выборе $_j$ следует руководствоваться в основном тем, чтобы ее величина была достаточно малой и общая структура кластеризации существенно не изменилась.

Если расщепление происходит на этом шаге, то надо перейти к шагу 2, в противном случае продолжать выполнение алгоритма.

Шаг 12. Вычисляются расстояния D_{ij} между всеми парами центров кластеров

$$D_{ij}= (Z_i, Z_j),\ i=1, 2,..., l_1-1,\ j=i+1,..., l_1. \qquad (2.38)$$

Шаг 13. Расстояния D_{ij} сравниваются с параметром $_c$. Те q расстояний, которые оказались меньше $_c$, ранжируются в порядке возрастания:

$$\left[D_{i_1,j_1}, D_{i_2,j_2},\ldots, D_{i_q,\ j_q}\right],$$

причем $D_{i_1,j_1}<D_{i_2,j_2},\ldots,D_{i_q,j_q}$, а q - максимальное число пар центров кластеров, которые можно объединить. Следующий шаг осуществляет процесс слияния кластеров.

Шаг 14. Каждое расстояние $D_{i_l,\ j_l}$ вычислено для определенной пары кластеров с центрами Z_{i_l} и Z_{j_l}. К этим парам в последовательности, соответствующей увеличению расстояния между центрами, применяется процедура слияния, осуществляемая на основе следующего правила.

Кластеры с центрами Z_{i_l} и Z_{j_l}, $l=1, 2,..., q$ объединяются, причем новый центр кластера определяется по формуле

$$Z_l^{\square} = \frac{1}{m_{il} + m_{jl}} \left[m_{il}\left(Z_{il}\right) + m_{jl}\left(Z_{jl}\right) \right]. \qquad (2.39)$$

Центры кластеров Z_{i_l} и Z_{j_l} ликвидируются и значение l_l уменьшается на 1.

Отметим, что допускается только попарное слияние кластеров и центр полученного в результате кластера рассчитывается, исходя из позиций, занимаемых центрами объединяемых кластеров и взятых с весами, определяемыми количеством объектов в соответствующем кластере. В процессе объединения центр нового кластера получится как истинное среднее сливаемых кластеров.

Шаг 15. Если текущий цикл итерации – последний, то выполнение алгоритма прекращается. В противном случае следует возвратиться либо к шагу 1 (если нужно поменять параметры алгоритма), либо к шагу 2, если управляющие параметры алгоритма остаются неизменными.

Пример. Выявление кластеров

Хотя алгоритм ИСОМАД не очень подходит для ручных вычислений, принцип его работы можно проиллюстрировать на простом примере.

Рассмотрим выборку, образы которой размещены так, как изображено на рис. 2.7.

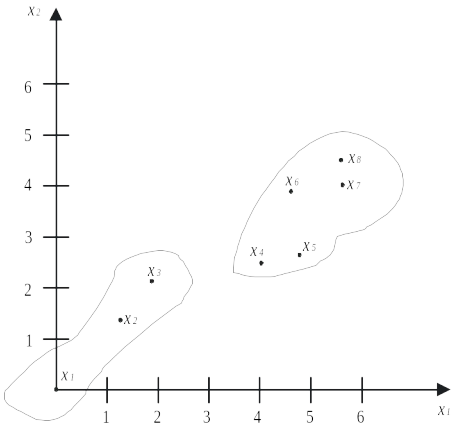

Рис. 2.7. Выборка образов, использованная для иллюстрации работыалгоритма ИСОМАД

В данном случае $N=8$ и $n=2$. В качестве начальных условий задаем $N_C=1, z_1=(0,0)'$ и следующие значения параметров процесса кластеризации:

Шаг 1.

$k=2, \theta_N=1, \theta_S=1, \theta_C=4; L=0, I=4.$

Если всякая априорная информация об анализируемых данных отсутствует, эти параметры выбираются произвольным образом и затем корректируются от итерации к итерации.

Шаг 2.

Так как задан только один центр кластера, то

$S_1=\{x_1, x_2, \ldots, x_8\}$ и $N_1=8$.

Шаг 3.

Поскольку $N_1>\theta_N$, ни одно подмножество не ликвидируется.

Шаг 4.

Корректируется положение центра кластера:

$z_1=1/N_1 \sum_{x \in S_1} x=\begin{pmatrix}3,38\\2,75\end{pmatrix}$.

Шаг 5.

Вычисляется расстояние $\bar{D}_j$:

$\bar{D}_1=1/N_1 \sum_{x \in S_1} \|x-z_1\|=2,26$.

Шаг 6.

Вычисляется расстояние $\bar{D}$:

$\bar{D}=\bar{D}_1=2,26$.

Шаг 7.

Поскольку данный цикл итерации - не последний и $N_C=K/2$, осуществляется переход к шагу 8.

Шаг 8.

Для подмножества S_1 вычисляется вектор среднеквадратичного отклонения:

$\delta_1=\begin{pmatrix}1,99\\1,56\end{pmatrix}$.

Шаг 9

Максимальная компонента вектора δ_1 равна 1,99, следовательно, $\delta_{1\max}=1,99$.

Шаг 10.

Поскольку $\delta_{1\max}>\theta_S$ и $N_C=K/2$, кластер с центром z_1 расщепляется на два новых кластера. Следуя процедуре, предусмотренной шагом 10, выбираем $\gamma_j=0{,}5$, $\delta_{j\max}\approx 1{,}0$. При этом

$z_1^+=\begin{pmatrix}4,38\\2,75\end{pmatrix}, z_1^-=\begin{pmatrix}2,38\\2,75\end{pmatrix}$.

Для удобства записи будем называть центры этих кластеров z_1 и z_2 соответственно. Значение N_C увеличивается на 1; переход к шагу 2.

Шаг 2.

Подмножества образов имеют теперь следующий вид:

$S_1=\{x_4, x_5, x_6, x_7, x_8\}, S_2=\{x_1, x_2, x_3\}$ и $N_1=5, N_2=3$.

Шаг 3.

Поскольку обе величины - и N_1, и N_2 - больше θ_N, ни одно подмножество не ликвидируется.

Шаг 4.

Корректируется положение центров кластеров:

$$z_1 = 1/N_1 \sum_{x \in S_1} x = \begin{pmatrix} 4,80 \\ 3,80 \end{pmatrix}, z_2 = 1/N_2 \sum_{x \in S_2} x = \begin{pmatrix} 1,00 \\ 1,00 \end{pmatrix}.$$

Шаг 5.

Вычисляется расстояние $\bar{D}_j, j = 1,2$:

$$\bar{D}_1 = 1/N_1 \sum_{x \in S_1} \|x - z_1\| = 0,80,$$

$$\bar{D}_2 = 1/N_2 \sum_{x \in S_2} \|x - z_2\| = 0,94$$

Шаг 6

Вычисляется расстояние $\bar{D}$:

$$\bar{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \bar{D}_j = \frac{1}{8} \sum_{j=1}^{2} N_j \bar{D}_j = 0,85.$$

Шаг 7

Поскольку данная итерация имеет четный порядковый номер, условие (в) шага 7 выполняется. Поэтому следует перейти к шагу 11.

Шаг 11

Вычисление расстояний между парами центров кластеров:

$$D_{12} = \|z_1 - z_2\| = 4,72.$$

Шаг 12

Величина расстояния D_{12} сопоставляется с параметром θ_C. В данном случае $D_{12} > \theta_C$.

Шаг 13

Результаты шага 12 показывают, что объединение кластеров невозможно.

Шаг 14

Поскольку данный цикл итерации - не последний, необходимо принять решение: вносить или не вносить изменения в параметры процесса кластеризации. Так как в данном (простом) случае: 1) число выделенных кластеров соответствует заданному; 2) расстояние между нами больше среднего разброса, характеризуемого среднеквадратичными отклонениями, и 3) каждый кластер содержит существенную часть общего количества выборочных образов, то делается вывод о том, что локализация центров кластеров правильно отражает специфику анализируемых данных. Следовательно, переходим к шагу 2.

Шаги 2- 6 дают те же результаты, что и в предыдущем цикле итерации.

Шаг 7

Ни одно из условий, проверяемых при реализации данного шага, не выполняется. Поэтому переходим к шагу 8.

Шаг 8

Для множеств $S_1 = \{x_4, x_5, x_6, x_7, x_8\}, S_2 = \{x_1, x_2, x_3\}$

$$\delta_1 = \begin{pmatrix} 0,75 \\ 0,75 \end{pmatrix}, \delta_2 = \begin{pmatrix} 0,82 \\ 0,82 \end{pmatrix}.$$

Шаг 9

В данном случае $\delta_{1\max}=0,75$ и $\delta_{2\max}=0,82$.

Шаг 10

Условия расщепления кластеров не выполняются. Следовательно, переходим к шагу 11.

Шаг 11

Полученный результат идентичен результату последнего цикла итерации
$D_{12}=\|z_1-z_2\|=4,72$.

Шаг 12

Полученный результат идентичен результату последнего цикла итерации.

Шаг 13

Полученный результат идентичен результату последнего цикла итерации.

Шаг 14

На данном цикле итерации не были получены новые результаты, за исключением изменения векторов среднеквадратичного отклонения. Поэтому переходим к шагу 2.

Шаги 2- 6 дают те же результаты, что и в предыдущем цикле итерации.

Шаг 7

Поскольку данный цикл итерации - последний, задаем $\theta_C=0$ и переходим к шагу 11.

Шаг 11

Как и раньше,
$D_{12}=\|z_1-z_2\|=4,72$.

Шаг 12

Полученный результат идентичен результатам последнего цикла итерации.

Шаг 13

Результаты шага 12 показывают, что объединение кластеров невозможно.

Шаг 14

Поскольку данный цикл итерации - последний, выполнение алгоритма заканчивается.