Кукса П.П.

Московский Государственный Технический Университет им. Н.Э. Баумана

E-mail: kouxa@online.ru
WWW: http://www.geocities.com/pkouxa

АНАЛИЗ АЛГОРИТМА НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ

Кластер-анализ составляет основу многих методов информационного моделирования и классификации. Целью кластеризации является получение компактного представления исходного набора данных с помощью кластеров. Методами нечеткого кластер-анализа можно индуцировать правила, то есть получить лингвистическое описание исследуемой системы, исходя из данных наблюдений (входо-выходной выборки). Индуцированный набор правил может затем модифицироваться экспертом. Одним из широко применяемых методов кластеризации является метод нечетких *c*-средних (*FCM*), сводящий кластеризацию к решению задачи структурного синтеза [1], т.е. к нахождению некоторого варианта структуры разбиения исходного множества данных на совокупность нечетких подмножеств. Целевой функцией задачи является:

$$J(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^{m}\cdot\left\|x_k-v_i\right\|^2, \qquad (1)$$

где $u_{ik}\in[0,1]$ - степень принадлежности x_k к c_i ; $m\in(1,\infty)$ -показатель нечеткости; $c\in\{2,3,\ldots,n-1\}$ -количество кластеров; n - мощность X ; $x_k\in\Re^p$ - k-е наблюдение (измерение) (p -мерный вектор); $v_i\in\Re^p$ -центр i-го кластера (p -мерный вектор); $U=[u_{ik}],i=\overline{1,c},k=\overline{1,n}$ - $c\times n$ -матрица разбиения; $V=[v_1,\ldots,v_c]^T$ - $c\times p$ - матрица прототипов (множество c центров v_i).

Формальная постановка задачи оптимальной кластеризации:

$$\text{Найти } U^*,V^*:J(U^*,V^*)\to\min\text{, при }\sum_{i=1}^{c}u_{ik}-1=0,\ 0\le u_{ik}\le 1,\ k=\overline{1,n} \qquad (2)$$

где U^*,V^* - оптимальной решение; $J(U^*,V^*)$ - оптимальное значение целевой функции; U,V - допустимые решения (варьируемые параметры).

Кластеризация сводится к оптимизации функции двух переменных при наличии ограничений в виде равенств. При фиксированном V задача эквивалентна минимизации функции Лагранжа:

$$L(U,\lambda)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^{m}\cdot\left\|x_k-v_i\right\|^2-\sum_{k=1}^{n}\lambda_k\Big(\sum_{i=1}^{c}u_{ik}-1\Big), \qquad (3)$$

где λ_k -множитель Лагранжа. При фиксированном U задача сводится к определению V из системы равенств:

$$\frac{\partial J(U,V)}{\partial v_i}=-\sum_{k=1}^{n}2\cdot(u_{ik})^{m}(x_k-v_i)=0,\ i=\overline{1,c} \qquad (4)$$

Из (3) и (4) можно получить следующие формулы для расчета v_i и u_{ik} :

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{\frac{2}{m-1}}}, \qquad i = \overline{1,c},\ k = \overline{1,n} \tag{5}$$

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \cdot x_k}{\sum_{k=1}^{n} (u_{ik})^m}, \quad i = \overline{1,c} \tag{6}$$

Предлагается [2] решать задачу нахождения решения U и V, целевая функция которого принимает экстремальное (минимальное) значение, итерационно. Схема классического варианта (первого варианта) алгоритма нечетких *c* - средних приведена на рис. 1. (двухшаговый алгоритм) При минимизации $J(U,V)$ в каждой итерации вычисляются U и V в соответствии с (5) и (6). Можно заметить, что для нахождения $\min(J(U,V)$ порядок вычисления U и V не имеет принципиального значения, поэтому возможна реорганизация процесса вычислений: можно одновременно вычислять U и V в едином цикле. Преобразование процесса вычислений U и V приводит к модифицированному алгоритму нечетких *c* – средних (рис. 2): алгоритм опять же представляет итерационный по $x_k, k = \overline{1,n}$ процесс, но фазы вычисления U и V переплетаются друг с другом.

Вычисление *V* по (6) (шаг 1) | Вычисление *U* по (5) (шаг 2) | Алгоритмическая модель модифицированного алгоритма *FCM* | Схема алгоритма *FCM*

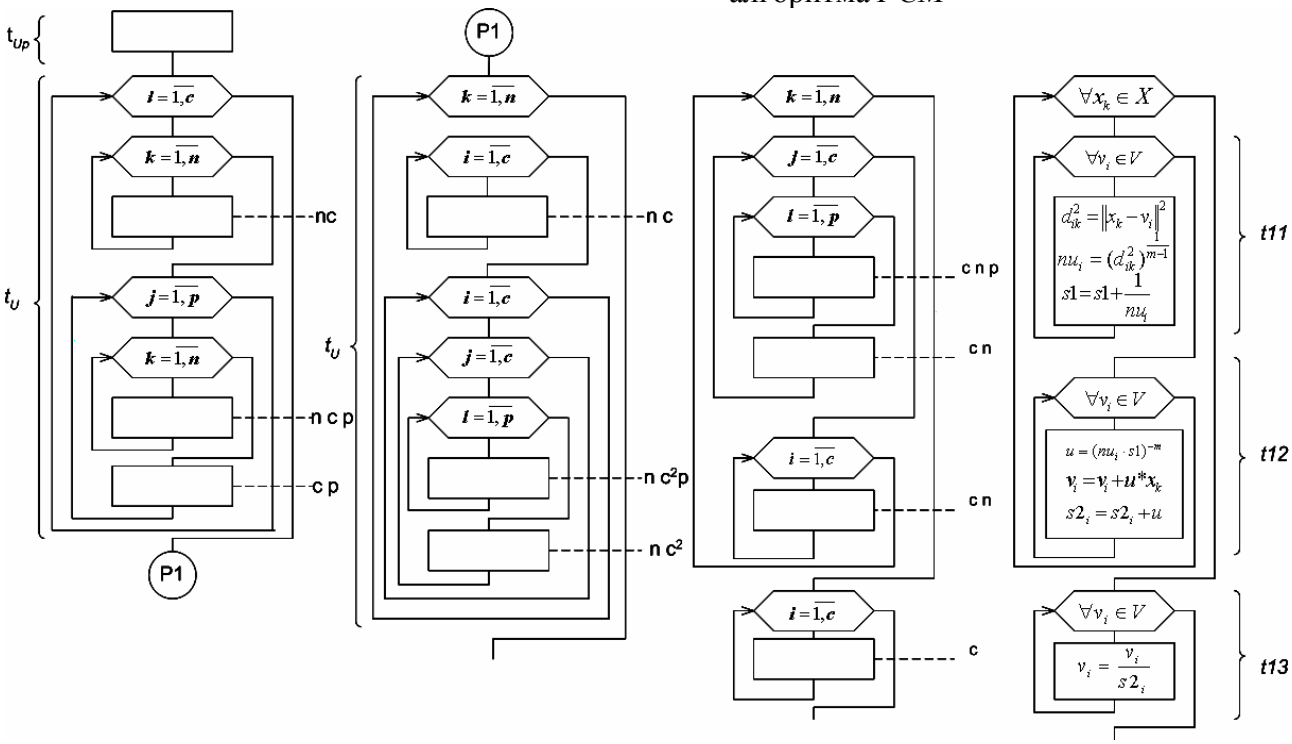

Рис. 1. Схема алгоритма FCM

Рис 2. Реорганизация алгоритма *FCM*

Рассмотрим организацию итерации. Итерация состоит из трех фаз: первые две многократно повторяются, последняя выполняется однократно.

Фаза 1(цикл от *i=1* до *c*): вычисляются $d_{ik}^2, i = \overline{1,c}$, где $d_{ik}^2 = \|x_k - v_i\|^2$; и накапливается сумма $s1 = \sum_{i=1}^{c} \left(\frac{1}{d_{ik}^2} \right)^{\frac{1}{m-1}}$. Фаза 2(цикл от *i=1* до *c*): определяются $u_{ik}^m, i = \overline{1,c}$ как $(d_{ik}^2 \cdot s1)^{-m}$;

накапливается сумма $s2_i = s2_i + u_{ik}^m$, $s3_i = s3_i + u_{ik}^m \cdot x_k$ и $J = \sum_{i=1}^{c} u_{ik}^m \cdot d_{ik}^2$. Фазы 1,2 повторяются от *k=1* до *n*. Фаза 3(цикл от *i=1* до *c*): определяются v_i как $\frac{s3_i}{s2_i}$. Фаза 3 следует после завершения цикла от *k=1* до *n*.

Оценка временной сложности реализаций алгоритма.

А. Первый вариант алгоритма

Вычисление нормы $\|x_k - v_i\|^2$: $t_{\|\|} = 4 + p \cdot (4 + t_+ + 2t_{[][]} + t_= + t_* + t_+ + t_=) = 4 + 84 \cdot p$

Вычисление V: $t_V = 4 + c(4 + t_{denom} + t_{nom})$,

где $t_{denom} = 4 + n(4 + t_{pow} + t_{[][]} + t_+ + t_=) = 4 + n \cdot (34 + t_{pow})$, t_{pow} - сложность возведения в степень; $t_{nom} = 4 + p(4 + (4 + n(4 + t_{pow} + t_+ + t_{[][]} + t_=)) + t_{[][]} + t_/ + t_=) = 4 + 64p + np(34 + t_{pow})$

$t_V = 4 + 12c + 64cp + ncp(34 + t_{pow}) + nc(34 + t_{pow})$

Вычисление U: $t_U = 4 + n(4 + t_{d2ik} + t_{uik})$, где $t_{d2ik} = 4 + c(4 + t_{\|\|} + t_{[]} + t_=) = 4 + 12c + 84cp$;

$t_{uik} = 4 + c(4 + t_{[][]} + t_/ + t_* + t_{pow} + 4 + c(4 + t_{\|\|} + t_= + t_= + t_+ + t_/ + t_* + t_{pow})) =$

$= 4 + c(82 + t_{pow}) + c^2 62 + 84c^2 p + c^2 t_{pow}$

Сложность одной итерации алгоритма:

$$t1 = t_U + t_V + nc(t_= + t_{[][]}) = \\ = 8 + 12n + 12c + 64cp + nc(156 + 2t_{pow}) + ncp(118 + t_{pow}) + nc^2(62 + t_{pow}) + (nc^2 p)84 \quad (7)$$

Б. Второй вариант алгоритма

$t11 = 4 + c(4 + t_{\|\|} + t_= + t_{[]} + t_{[]} + t_= + t_{pow} + t_{[]} + t_+ + t_/ + t_=) = 4 + 50c + 84cp + c \cdot t_{pow}$

$t12 = 4 + (86 + 2t_{pow})c + 54cp$; $t13 = 4 + c(4 + t_{[]} + t_{[]} + t_/) = 4 + 84c$

Сложность одной итерации алгоритма:

$$t1 = 4 + n(4 + t11 + t12) + t13 = 8 + 12n + 84c + (136 + 3t_{pow})nc + 138 \cdot (ncp) \quad (8)$$

Сравнение А и Б:

$$\Delta_{12} = nc^2 p * 84 + nc^2(t_{pow} + 62) + ncp(t_{pow} - 20) - nc(t_{pow} - 20) + 64cp - 72c \quad (9)$$

Оценка емкостной сложности

Структуры данных и их емкостная сложность: $X = M_{pn}(\Re)(p \cdot n \cdot sz)$;

$V = M_{cp}(\Re)$ $(p \cdot c \cdot sz)$; $U = M_{cn}(\Re)$ ($2 \cdot c \cdot n \cdot sz$ - для первого варианта; $c \cdot n \cdot sz$ - для второго варианта); $d2 = M_{1c}(\Re)(c \cdot sz)$; $d2ik = M_{1c}(\Re)(c \cdot sz)$ (только для второго варианта).

Следовательно, общая емкостная сложность реализации первого и второго варианта составляет $C1 = sz \cdot (n \cdot (p + 2 \cdot c) + c(p + 1))$ и $C2 = sz \cdot (n(p + c) + c(p + 2)$, соответственно.

$$\Delta C12 = sz \cdot c \cdot (n - 1) \quad (10)$$

Реорганизация алгоритма приводит к снижению временной и емкостной сложности алгоритма (см. (9) и (10)). Оценка вычислительной сложности реализаций алгоритма по алгоритмической модели (рис. 1, 2) и асимптотическая оценка вычислительной сложности приведены в табл. 1.

Таблица 1.

Алгоритм	вычислительная сложность	асимптотическая оценка
Первый вариант	$n \cdot (c^2 p + c^2 + c) + n \cdot p \cdot c + nc + cp$	$O(nc^2 p)$
Второй вариант	$n \cdot c \cdot p + 2 \cdot c \cdot n + c$	$O(ncp)$

Литература

1. Овчинников В.А. Алгоритмизация комбинаторно-оптимизационных задач при проектировании ЭВМ и систем. – М.: Изд-во МГТУ им. Н.Э.Баумана, 2001 г. – 288 с.

2. J. C. Bezdek, Pattern Recognition with Fuzzy objective Function Algorithms, Plenum Press, 1981.

3. F. Hoppner, F. Klawonn Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. New York: Willey, 1999.