XXXIII Неделя науки СПбГПУ. Материалы межвузовской научно-технической конференции. Ч.VI: С.60-62, 2005.
© Санкт-Петербургский государственный политехнический университет, 2005.

УДК 65.9 (2)23; 22.18; 30.607

К.А.Криулькина (5 курс, ГУАП), Е.Г.Семенова, д.т.н., проф. ГУАП

КЛАСТЕРНЫЙ АНАЛИЗ ЭКСПЕРТНЫХ ОЦЕНОК АЛЬТЕРНАТИВНЫХ ВАРИАНТОВ

В настоящем сообщении рассматривается методика кластерного анализа альтернативных вариантов, представленных в виде экспертных оценок по качественным признакам.

Для совокупности *k* объектов $X = \{x_1,...,x_k\}$, оцениваемых *n* экспертами по критериям $Q_1, Q_2, ...Q_m$, имеющим шкалы качественных оценок $q_{ir}^{e_r}$, r =1,…,*m*; e_r=1,…,h_r, введем множество значений оценок $G = \{g_1,...,g_h\}$, элементы которого определим следующим образом:

$$g_1 = q_1^1,...,q_{h_1} = q_1^{h_1}, g_{h_1+1} = q_2^1,...,q_h = q_m^{h_m}, h = h_1 + h_2 + ... + h_m. \quad (1)$$

Каждый объект $x_i \in X$ может быть записан в виде:

$$x_i = \{n_i(g_1) * g_1,...n_i(g_h) * g_h\}, \quad (2)$$

где $n_i(g_j)$ равно числу экспертов, давших объекту x_i оценку $g_j = q_j^{e_j}$.

Совокупность объектов $X = \{x_1,...,x_k\}$ вида $x_i = \{n_i(g_j) * g_j\}$ образует мультимножество (ММ). Для мультимножеств справедливы операции объединения, пересечения, суммы, разности, симметрической разности и умножения на скаляр.

Задача кластеризации применительно к мультимножествам сводится к разбиению исходной совокупности объектов $X = \{x_i\}$ на несколько групп $\{X_1,...,X_k\}$ на основе сходства или различия их свойств $G = \{g_j\}$.

Выбор показателя близости, сходства или различия между объектами зависит от их физической или статистической природы. Для бинарных данных, описываемых в шкалах отношений или интервальных, а также для исследуемых объектов типа мультимножеств, пространства которых неевклидовы, более пригодны модели пространств, основанные на рассмотренных метриках [1-3]. За основу выбора модели классификации мультимножеств принимается естественное предположение о том, что показатель различия/сходства между объектами $x_a, x_b \in X$ и между группами объектов (кластерами) $X_p, X_q \in X$ должен быть одного и того же типа, например, определяться метриками вида $d(A,B)$ или $s(A,B)$.

Выражения для расстояний и мер сходства между кластерами мультимножеств будут иметь вид:

$$\begin{array}{l|l} d_0(X_p, X_q) = D_{pq} & s_1(X_p, X_q) = 1 - (D_{pq}/W) \\ d_2(X_p, X_q) = D_{pq}/M_{pq} & s_2(X_p, X_q) = I_{pq}/M_{pq} \\ d_1(X_p, X_q) = D_{pq}/W & s_3(X_p, X_q) = I_{pq}/W \end{array} \quad (3)$$

Здесь

$$I_{pq} = \sum_{j=1}^{h} w_j \min(c'_{pj}, c'_{qj}) \quad M_{pq} = \sum_{j=1}^{h} w_j \max(c'_{pj}, c'_{qj})$$
$$D_{pq} = \sum_{j=1}^{h} w_j \mid c'_{pj} - c'_{qj} \mid \qquad W = \sum_{j=1}^{h} w_j \tag{4}$$

Элементы матриц c'_{tj} определяются в зависимости от способа группирования объектов.

Для сравнительного исследования различных методов кластерного анализа мультимножеств были использованы серии модельных экспериментов. В качестве исходных данных приняты совокупности из 20объектов (k=20) , оцениваемых 10 экспертами (n=10) по 4 критериям (m=4). Шкалы по всем критериям были разбиты на 5 оценок, т.е. $h = 4 \times 5 = 20.$

Исследование проводилось на двух различных вариантов совокупностей объектов:

- с признаками, имеющими случайные оценки с равномерным распределением;
- с ярко выраженными признаками, на значения которых наложен белый шум.

Разработанные и апробированные алгоритмы иерархического и неиерархического анализа по форме имеют много общего. Достаточно отметить, что при $K = 1,..., N$ неиерархический алгоритм переходит в иерархический и наоборот.

Принципиальное же отличие состоит в том, что при иерархическом подходе за основу берется вся совокупность объектов, входящих в кластер, т.е. он представлен либо верхней «огибающей» (объединение), либо комбинацией (сумма, взвешенная сумма) входящих в него объектов. В этом случае рассматриваются все свойства группы объектов, но почти никак не затрагивается «форма» образуемых кластеров.

При неиерархическом подходе замена кластера его «центром» сопровождается частичной потерей информации. Но при этом в основу кладется именно «форма» кластера – кучность образующих его объектов.

Именно по этой причине нельзя рекомендовать использование только одного подхода в случае наличия или отсутствия начального представления о структуре системы. Только совместный анализ может дать ЛПР представление о структуре анализируемого множества. В частности, при тестовых экспериментах иерархические методы показали, что система состоит из трех групп, а не из четырех, как предполагалось изначально.

В то же время неиерархический подход позволяет обнаружить подкатегории среди образующихся групп. Сильное расхождение результатов для случайного набора данных позволяет говорить об отсутствии структуры в начальном множестве.

Таким образом, полученные результаты позволяют сформулировать ряд рекомендаций. Наиболее адекватные результаты при кластерном анализе экспертных оценок альтернативных вариантов соответствуют совместному использованию линейной комбинации для слияния объектов в кластер и метрики d_1 для вычисления расстояний. Иерархический подход к анализу множеств более устойчив к начальным условиям и может быть рекомендован к использованию даже тогда, когда количество кластеров заранее известно. В методах неиерархического анализа можно рекомендовать использование метрики d_1, как более устойчивой, и проведение возможно большего числа испытаний с различными начальными условиями для нахождения по возможности большего числа локальных экстремумов.

ЛИТЕРАТУРА:

1. Semenova E.G. Algorithms of cluster analysis in an assessment of qualitative alternatives. International conference «Instrumentation in Ecology and Human Safety». S.Petersburg, 2002.
2. Кондратенков В.А., Исаев С.А., Ипатко И.В. Вопросы теории надежности технических систем. – Смоленск: Изд.во Смоленского университета, 1998. – 169 с.

3. Дубов Ю.А. Многокритериальные модели формирования и выбора вариантов системы. – СГТУ, Саратов, 2000. – 295 с.