Коллективные решения задачи кластерного анализа с помощью гиперграфов

Шмаков А. С.
ashmak@mail.ru

Исследуется выборка $S = x_1, \ldots, x_m$, заданная таблицей признаковых описаний

$$J_m(S) = \left\| \begin{matrix} a_{11} & \ldots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{md} \end{matrix} \right\| \tag{1}$$

на множестве вещественно-значимых признаков, где d — число признаков, а m — число объектов.

Пусть в результате работы некоторого алгоритма A исходная выборка разбита на l кластеров $K_1, \ldots, K_l$.

Определение 1. *Информационной матрицей для задачи кластеризации выборки (1) на l кластеров называется матрица $I = \|\alpha_{ij}\| \in R_{m\times l}$, такая что $\alpha_{ij} \in \{0, 1, \triangle\}$, причем $\alpha_{ij} = 1$, если $x_i \in K_j$; $\alpha_{ij} = 0$, если $x_i \notin K_j$; $\alpha_{ij} = \triangle$ соотвествует отказу от зачисления x_i в один из классов.*

В отличие от задач распознавания, где каждый кластер имеет свое априорное смысловое содержание, в задачах кластерного анализа кластеры могут нумероваться в произвольном порядке. Поэтому произвольная информационная матрица I, отличающаяся от I' лишь перестановкой столбцов, будет являться записью того же самого решения.

Пусть теперь рассматривается задача кластеризации исходной выбоки (1) на l кластеров коллективом из n алгоритмов $A^1, \ldots, A^n$, и, кроме того, пусть в результате анализа исследуемый выборки получены n матрицы оценок (информационных матриц) $I^1, \ldots, I^n$.

Определение 2. *Коллективной информационной матрицей $\hat{I} = \|\chi_{ij}\|$ для задачи кластеризации выборки (1) на l кластеров коллективом из n алгоритмов называется блочная матрица размера $m \times ln$, состовленная из информационных матриц $I^1, \ldots, I^n$, т.е. $\hat{I} = \|I_1 \cdots I_n\|$.*

На основании коллективной информационной матрицы можно ввести понятие гиперграфа сопряженности.

Определение 3. *Гиперграфом сопряженности G_H для задачи коллективной кластеризации называется гиперграф, для которого коллективная информационная матрица $\hat{I}$ является матрицей инцидентности.*

Для построения коллективного решения предлагается построить разбиения гиперграфа сопряженности на l узлов в следующем виде: пусть

дан гиперграф $G_H(V, E)$, где множество вершин $V = \{x_1, \ldots, x_n\}$ — объекты выборки, а множество ребер $E = \{e_j, j = 1, \ldots, ln\}$ — столбцы коллективной информационной матрицы. Вес вершин считается одинаковым и равным 1, а вес ребер задается множеством $\varphi_j = \sum_{i=1}^{m} \chi_{ij}$, $j = 1, \ldots, ln$, где χ_{ij} — элементы коллективной информационной матрицы.

Для решения поставленной задачи предлагается использовать известный алгоритм разделения гиперграфа — HMETIS. Упрощенно алгоритм работает следующим образом:

1. Coarsening phase — фаза стягивания или огрубления, на которой строится последовательность грубых приближений гиперграфа:
   - Edge coarsening — простейший способ группировки вершин состоит в выборе пары вершин, принадлежащих одному и тому же гиперребру.
   - Hyperedge coarsening — выбирается независимое множество гиперребер; вершины, принадлежащие одному гиперребру, объединяются.
2. Initial partitioning phase — фаза разделения, на которой наименьший граф подвергается декомпозиции: в качестве алгоритма деления может быть использовано спектральное деление, геометрическое деление или комбинаторное деление.
3. Refinement phase — фаза, на которой решение для наименьшего графа проецируется на следующий уровень и уточняется итерационным алгоритмом Kernighan-Lin.

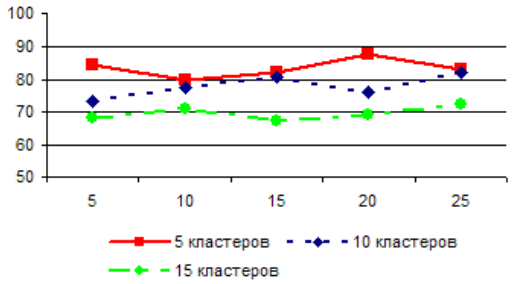
Таким образом, для решения задачи поиска локально оптимального решения по конечному набору кластеризаций предлагается следующий алгоритм:

1. Строим набор информационных матриц для каждого решения из коллектива $\{I_1, \ldots, I_n\}$;
2. Стром общую информационную матрицу для коллективного решения $\hat{I} = \|I_1 \cdots I_n\|$;
3. Строим гиперграф сопряженности H_G;
4. Используя алгоритм HMETIS, строим разделение гиперграфа H_G на l узлов;
5. Полученные в результате l подмножеств являются коллективным решением задачи.

Эксперименты с использованием гиперграфов проводились на выборках размера 500 объектов, состоящих из кластеров, которые имеют нормальные плотности распределения объектов по признакам. Были проведены испытания для выборок состоящих из $5, 10, 15$ кластеров. В качестве алгоритмов, участвующих в коллективах, использовались эври-

стические алгоритмы: k-внутригрупповых средних и алгоритм Форель. Количество решений для каждой задачи было 10. При исследованиях изучалась зависимость точности решения описываемым методом от возрастающей сложности решаемой задачи. Ниже приводится график зависимости точности кластеризации от числа признаков, описывающих выборку, при разном числе кластеров. На представленном ниже графике по оси Y отложено количество правильно классифицированных объектов в процентах, а по X — количество признаков, которые имела исследуемая выборка.

Работа выполнена при поддержке РФФИ, проект №05-07-90333, Целевой программы №14 Президиума РАН, Целевой программы №2 Отделения математических наук РАН.



Литература

[1] *Рязанов В. В.* О синтезе классифицирующих алгоритмов на конечных множествах алгоритмов классификации (таксономии).— ЖВМиМФ, 1982. — Т. 22, №2. — С. 429–440.

[2] *Рязанов В. В.* Комитетный синтез алгоритмов распознавания и классификации.— ЖВМиМФ, 1981. — Т. 21, №6. — С. 1533–1543.

[3] *Karypis G., Kumar V.* Multilevel k-way Hypergrap Partitioning. — VLSI Design. — Vol. 11, No. 3. — 2000, Pp. 285–300.

[4] *Karypis G., Selvakkumaran N.* Multi-Objective Hypergraph Partitioning Algorithms for Cut and Maximum Subdomain Degree Minimization. — IEEE Transactions on CAD, 2005.