Кластеризация больших массивов данных в смысле поиска “минимальных звезд”

Васильев Игорь Леонидович

Институт динамики систем и теории управления СО РАН

Иркутск, 2012

Постановка задачи

Кластерный анализ

Разбиение исходного множества объектов на подмножества схожих объектов (кластеров).

Дано

Множество объектов

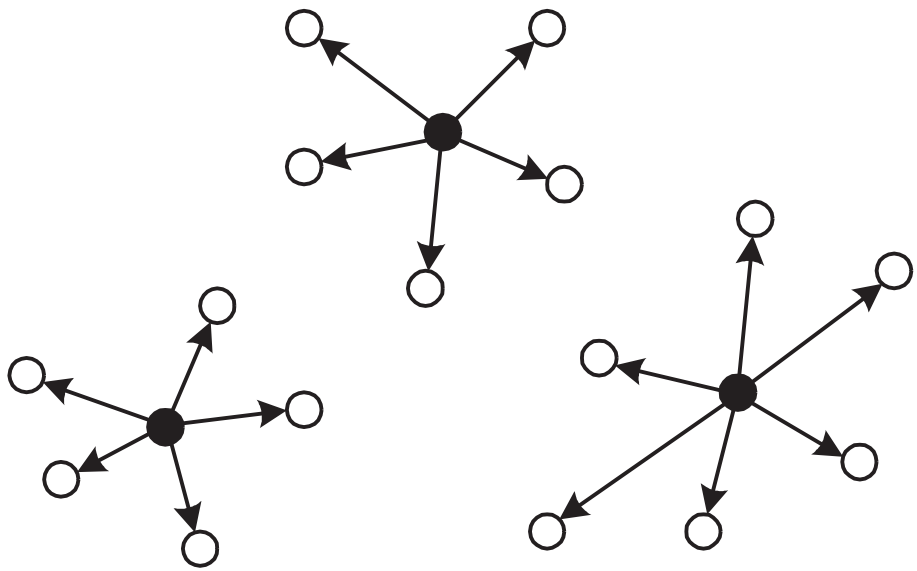
$$V = \{1, \ldots, m\}$$

$$a^u \in \mathbb{R}^n$$

$$d_{uv} = d(u, v) = \|a^u - a^v\|, \quad \forall u, v \in V$$

Постановка задачи

$G(V, A)$ – полный взвешанный диграф.
V – множество вершин,
$A = \{uv : u \in V, v \in V, u \neq v\}$ – множество дуг,
d_{uv} – веса дуг,
p – количество кластеров.
Необходимо найти p минимальных звезд, т.е. найти p вершин (медиан), минимизируя сумму весов дуг до остальных вершин.



Постановка задачи

Задача о p медиане в формулировке целочисленного программирования

$$y_i = \begin{cases} 1, \text{ если } i\text{-ая вершина медиана} \\ 0, \text{иначе} \end{cases}.$$

$$x_{ij} = \begin{cases} 1, \text{ если } j\text{-ая вершина к } i\text{-ой медиане} \\ 0, \text{иначе} \end{cases}.$$

$$Z^* = \min_{(x,y)} \sum_{u \in V} \sum_{v \in V} d_{uv} x_{uv} \tag{1}$$

$$\sum_{u \in V} x_{uv} + y_v = 1 \qquad \forall v \in V, \tag{2}$$

$$x_{uv} \leq y_u \qquad \forall u, v \in V, \tag{3}$$

$$\sum_{v \in V} y_v = p \tag{4}$$

$$y_u \in \{0, 1\} \qquad \forall v \in V, \tag{5}$$

$$x_{uv} \in \{0, 1\} \qquad \forall u, v \in V. \tag{6}$$

Поставленные задачи

- Разработан эвристический алгоритм, позволяющий решать задачи, возникающие в кластерном анализе с более чем со 100 тысячами объектов.
- Для повышения эффективности предложенного подхода планируется разработать параллельную реализацию алгоритма, которая должна позволить увеличь размерность рассматриваемых задач на порядок.