Медиана Кемени в задаче ранжирования и определения критической массы экспертов

Григорьев Алексей Дмитриевич

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Москва, 2021

Постановка задачи

Дано

$\mathfrak{D} = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x_i} \in \mathbb{R}^k$ — вектор экспертных оценок, n — число объектов, k — число экспертов.

Статистические предположения о данных

$\mathcal{R} = \{\mathbf{r}_j\}_{j=1}^k = \mathfrak{D}^T$, где $\mathbf{r}_j \in \mathbb{R}^n$ — вектор, задающий ранжирование объектов j-ым экспертом,
$\mathcal{R} \overset{i.i.d.}{\sim} \mathcal{D}$, где $\mathcal{D}$ — фиксированное распредление, с конечным мат.ожиданием: $\exists \mathbb{E}\mathcal{D}$.

Задачи

- Построить интегральный индикатор.
- Определить оптимальное число экспертов: построить процедуру принятия решения о достаточности текущего числа экспертов.

Метод Кемени-Янга (медиана Кемени)

Отношение частичного порядка

Для рейтинга $\mathbf{r}_l$, $\forall l = \overline{1, k}$ водится матрица попарных сравнений

$$\mathbf{A}^l = \begin{bmatrix} a^l_{11} & \dots & a^l_{1n} \\ \vdots & \ddots & \vdots \\ a^l_{n1} & \dots & a^l_{nn} \end{bmatrix}, \; a^l_{ij} = \begin{cases} 1, & r_{li} > r_{lj} \text{ (т.е. } x_{il} \succ x_{jl}) \\ 0, & \textbf{не сравнимы} \text{ или равны} \\ -1, & r_{li} < r_{lj} \text{ (т.е. } x_{il} \prec x_{jl}) \end{cases}$$

Пропуск объекта в рейтинге соответствует его несравнимости.

Расстояния на основе попарных сравнений

- Расстояние между парой рейтингов $\mathbf{r}_l$, $\mathbf{r}_m$:
$$d(\mathbf{r}_l, \mathbf{r}_m) = \sum_{i=1}^n \sum_{j=1}^n |a^l_{ij} - a^m_{ij}|$$
- Расстоние от рейтинга $\mathbf{r}$ до мн-ва рейтингов $\mathcal{R} = \{\mathbf{r}_l\}_{l=1}^k$:
$$d(\mathbf{r}, \mathcal{R}) = \sum_{l=1}^k d(\mathbf{r}, \mathbf{r}_l)$$

Метод Кемени-Янга (медиана Кемени)

Медиана Кемени

Медиана Кемени для выборки рейтингов $\mathcal{R}$ определяется как решение задачи оптимизации:

$$\mathbf{r}^* = \arg\min_{\mathbf{r}} d(\mathbf{r}, \mathcal{R}), \text{ где}$$

$$\mathbf{r} \; - \text{перестановка ранжирования.}$$

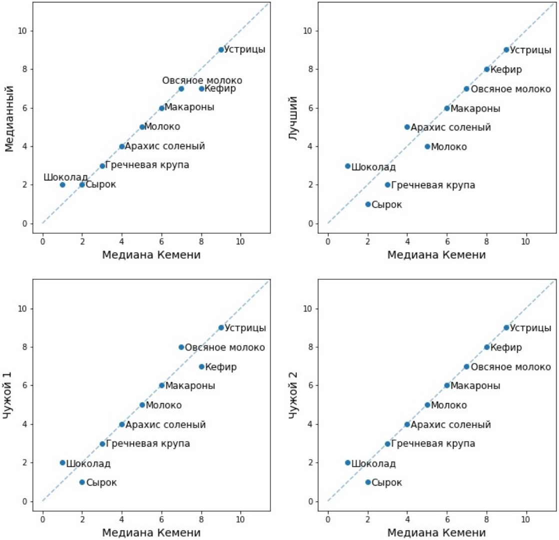
Интегральный индикатор $\mathcal{I}_i$ объекта i через медиану Кемени

$$\mathcal{I}_i = -r_i^*$$

Свойства

- *частичный* порядок $\Rightarrow$ $\exists$ устойчивость к ***пропускам***;
- NP-hard $\Rightarrow$ алгоритм полного перебора всех перестановок;

Результаты: сравнения рейтингов



Отказ от принятия решения

Предложения

- Метод устойчив к пропускам, медиана Кемени существует для произвольной выборки рейтингов.
  Чрезмерность числа пропусков определяется спецификой конкретной задачи и интерпретацией результата.
- Альтернативно может быть рассмотрен ансамбль других методов построения рейтинга и произведено сравнение с предложенным методом на предмет согласованности.
  Критерий сравнения — ранговое расстояние Кендалл-тау.

Оптимальное число экспертов

Принятие решения при пошаговом добавлении экспертов

- "Лидером" на шаге k назовем перестановку, задаваемую медианой Кемени для текущей выборки $\mathcal{R}_k = \{\mathbf{r}_l\}_{l=1}^{k}$.
- **Критерий останова:** отсутвие смены "лидера" на протяжении последних N шагов (N — гиперпараметр).

Допущения *(условия применимости)*

- $\mathcal{R}_k \overset{i.i.d.}{\sim} \mathcal{D} : \exists \mathbb{E}\mathcal{D}$;
- $\exists!$ медиана Кемени;

Сходимость

В силу ЦПТ расстояние от некоторой перестановки до выборки, нормированное на число экспертов, сходится к мат.ожиданию.

Оптимальное число экспертов: эксперимент

Данные (согласованные "зашумленные" эксперты)

$$\mathcal{R}_k \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}),\ \boldsymbol{\mu} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix}^T,\ \sigma = 4$$