

Парето-ранжирование с предпочтениями на признаках

Панченко Святослав

Московский Физико-Технический Институт
Факультет Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Москва,
2021 г.

Построение рейтинга продуктов питания

Задача

Построить рейтинг для совокупности различных продуктов питания. Каждый вид еды характеризуется набором числовых признаков - калорийность, цена, содержание белков, жиров, углеводов (в расчёте на 100г) и набором экспертных оценок качества.

Требования к модели построения рейтинга

- устойчивость относительно добавления новых объектов;
- полный порядок на объектах;
- учёт пропусков в значениях признаков и экспертных оценок.

Идея

Строить рейтинг предлагается с помощью Парето-ранжирования с предпочтениями на признаках.

Выборка

$$\mathcal{D} = \{x_i\}, i \in \mathcal{I} = \{1, \dots, m\},$$

$$x = [\chi_1, \dots, \chi_j, \dots, \chi_d], j \in \mathcal{J} = \{1, \dots, d\},$$

$\chi_j \in \mathbb{L}_j = \{l_1, \dots, l_{k_j}\}$ — частично упорядоченное множество.

Задача

Построить интегральный индикатор u для заданного множества объектов \mathcal{D} .

Определение n -отношения

Мы говорим, что объект $x_i = [x_{i1}, \dots, x_{id}]$ n -доминирует объект $x_j = [x_{j1}, \dots, x_{jd}]$,

$$\text{или } x_i \succ_n x_j,$$

если $x_{ik} \succeq x_{jk}$ для всех $k = 1, \dots, d$, но $x_i \neq x_j$.

Определение \tilde{n} -отношения

Мы говорим, что объект $x_i = [x_{i1}, \dots, x_{ir}, \dots, x_{it}, \dots, x_{id}]$ \tilde{n} -доминирует объект x_j с предпочтением на признаках $r \succ t$,

$$\text{если } x_i \succ_n x_j,$$

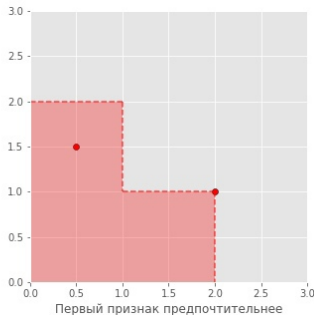
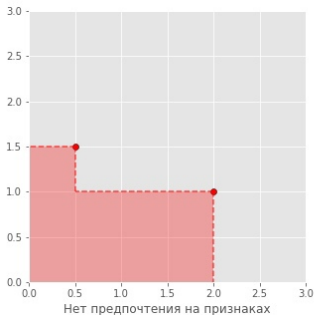
$$\text{или } x_i^{tr} \succ_n x_j,$$

где $x_i^{tr} = [x_{i1}, \dots, x_{it}, \dots, x_{ir}, \dots, x_{id}]$ - вспомогательный объект с измененным порядком следования признаков r и t .

Парето-оптимальный фронт

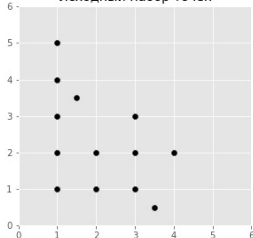
Набор объектов x_i , $i \in \mathcal{I}$, называется Парето-оптимальным фронтом POF_n , если для любого $x_i \in POF_n$ не существует объекта x , который бы его доминировал: $x \succ_n x_i$ (или $x \succ_{\tilde{n}} x_i$).

Парето-оптимальный фронт и его зона доминирования для пары точек в случае отсутствия и присутствия предпочтения на признаках

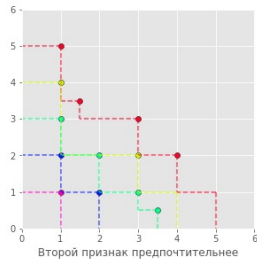
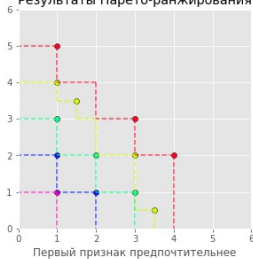
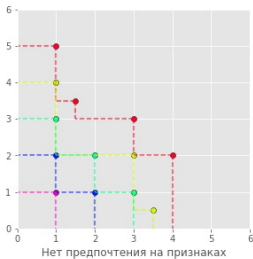


Парето-ранжирование

Исходный набор точек



Результаты Парето-ранжирования



Комбинирование рейтингов

Проблема

Построение рейтинга указанным методом в пространстве высокой размерности приводит к выделению очень небольшого числа фронтов, поскольку объекты, как правило, несравнимы между собой.

Подзадача

Предложить способ комбинирования рейтингов. В качестве комбинируемых рейтингов могут выступать промежуточные рейтинги или оценки экспертов.

Идея

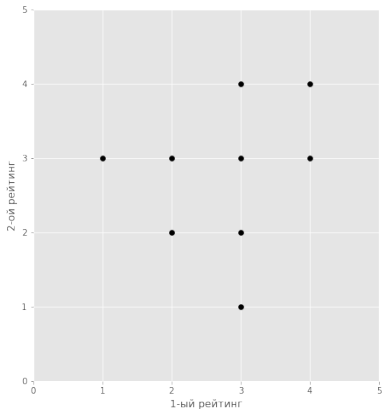
Позиции в промежуточных рейтингах рассматриваются в качестве нового признакового описания для построения нового рейтинга. Промежуточные рейтинги формируются на основе разбиения множества признаков на подмножества.

Предположим, что построены два промежуточных рейтинга φ_1 и φ_2 . В таком случае каждому объекту x соответствуют его позиции в этих рейтингах:

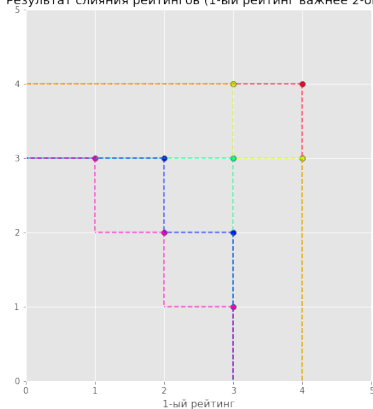
$$y_i = \varphi_i(x), \quad i = 1, 2.$$

Тогда вектор $[y_1, y_2]^T$ выступает в роли нового признакового описания объекта x ; на основе совокупности таких описаний строится новый рейтинг.

Слияние рейтингов на основе Парето-ранжирования, иллюстрация



Результат слияния рейтингов (1-ый рейтинг важнее 2-ого)



Исходное множество признаков разбито на смысловые группы, для каждой из которых строится отдельный рейтинг. Группы упорядочены по убыванию предпочтения для построения итогового рейтинга.

- Экспертные оценки - 14 признаков (на момент составления рейтинга):
 - заполнение пропусков медианой;
- Соотношение калорийность-цена - 2 признака:
 - сравнение по цене происходит по принципу "чем меньше, тем лучше";
 - показатель цены считаем важнее.
- Соотношение белки-жиры-углеводы - 3 признака:
 - сравнение по показателю содержания жиров происходит по принципу "чем меньше, тем лучше";
 - показатель содержания жиров считаем наименее важным.

Проблема 1: размерность

Вместо построения Парето-ранжирования на основе совокупности экспертных оценок предлагается использовать агрегированную оценку - *медиану Кемени*.

Медиана Кемени

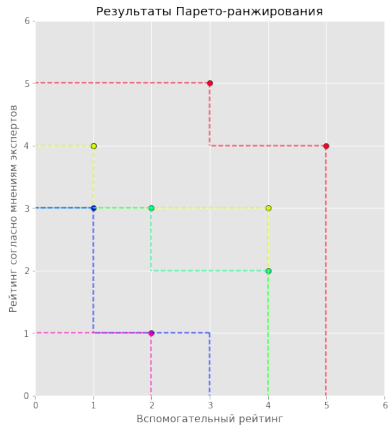
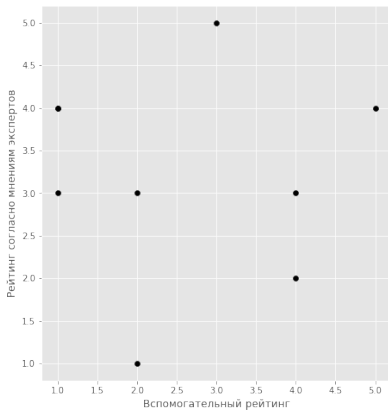
Пусть имеется k ранжирований на n объектах, заданных своими матрицами попарных предпочтений $A_u = \|a_{i,j}^{(u)}\|_{i,j=1}^n$, $u = 1, \dots, k$, $a_{i,j} \in \{-1, 0, 1\}$ для всех $i, j = 1, \dots, n$. Медиана Кемени - это ранжирование, задаваемое матрицей, наименее удалённой от остальных:

$$A^* = \arg \min \sum_{u=1}^k d(A, A_u) = \arg \min \sum_{u=1}^k \sum_{i,j=1}^n |a_{i,j} - a_{i,j}^{(u)}|$$

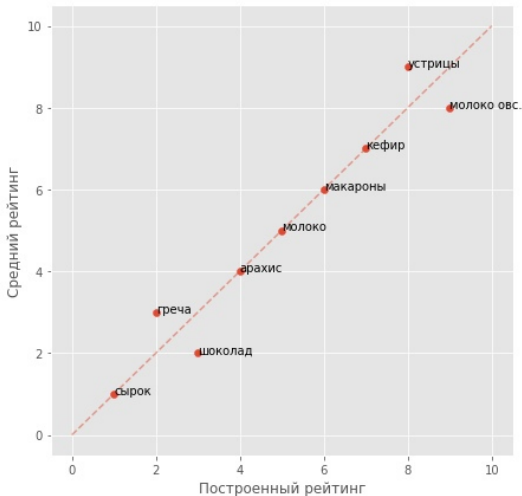
Статистические предположения о данных:

Для *i.i.d.* ранжирований A_1, \dots, A_k справедливо, что медиана Кемени $A^* \xrightarrow{P} \mathbb{E}A_u$ при $k \rightarrow \infty$.

Проблема 2: полный порядок



Сравнение полученного рейтинга со средним



- *Medvednikova M. M., Kuznetsov M. P., Strijov V. V. (2015)* 'Ordinal classification using Pareto fronts', Expert Systems with Applications.
- *Mironenkov A.A. (2019)* 'Hierarchical Pareto Classification of the Russian Regions by the Population's Quality of Life Indicators', Economic and Social Changes: Facts, Trends, Forecast.
- *С.Д. Двоенко, Д.О. Пшеничный, А.В. Попов (2017)* 'Групповое Ранжирование на Основе Медианы Кемени с Метрическими Свойствами', Известия ТулГУ, Технические науки, Вып. 10.