Построение интегрального индикатора с исследованием обработки пропусков

Вайсер Кирилл Олегович

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Москва
2020 г

Задача построения интегрального индикатора

Цель

Предложить способ построения интегрального индикатора для составления рейтинга продуктов.

Решаемая проблема

Описания объектов могут содержать пропущенные значений некоторых показателей. Так как интегральный индикатор служит средством ранжирования объектов, неясно как интерпретировать такие пропуски. Необходимо предложить метод обработки пропущенных значения при подсчете расстояния между объектами.

Идея

Метод решения

Предлагаемый метод заключается в следующем:

1. Ввести расстояние Минковского на пространстве объектов.
2. Определить идеал объекта.
3. Для каждого объекта определить значение расстояния до идеального.
4. Построить рейтинг на основе полученных расстояний.

Расстояние Минковского

Расстояние Минковского в евклидовом пространстве зависит от параметра p и определяется как

$$\rho(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{1}$$

При разных значениях p линии уровня выглядят по разному

1. $p < 1$
   При $p \to 0$ больший вклад вносят малые отклонения в большом числе признаков. Соответственно соседними будут объекты, которые имеют минимальное число отклонений в показателях.
2. $p \approx 2$
   При значениях $p > 1$ линии уровня напоминают привычную евклидову окружность. Из этого следует, что соседними будут объекты, равномерно близкие по всем показателям.
3. $p \to \infty$
   При значениях $p \to \infty$ линии уровня стремятся превратиться в квадрат. В таком случае соседними будут объекты, максимальное значения отклонения показателей которых минимально.

Выбор идеала

Пусть дана выборка объектов $\{x^i\}_{i=1}^n$, $x^i \in \mathbb{R}^d$ и веса показателей $\{w_j\}_{j=1}^d$. Идеальный объект x^* можно выбрать несколькими способами.

1. Объект с нулевыми или максимально возможными показателями.
2. Среднее между объектами:

$$x^* = \phi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(x^i)\right),$$

   где ϕ —непрерывная строго монотонная функция, индуцирующая тип среднего. Например в одномерном случае при $\phi(x) = x$ будет индуцировано среднее арифметическое.
3. Лучший из объектов

$$x^* = \text{argmax}_{x \in \{x_i\}} M(x, \{w_j\}),$$

   где M — функция, индуцирующая тип максимума. Например при $M(x, \{w_j\}) = \max_j(w_j)x_{\text{argmax}_j(w_j)}$ идеальным будет объект, обладающим наибольшим значением показателя, имеющего максимальный вес.

Обработка пропусков

Пусть m обозначает индекс пропущенного значения. Предлагаются следующие подходы для обработки пропусков.

1. Заполнить значение на основе статистик выборки

$$x_m = T(x_i, m)$$

   Например, можно заполнить пропуски средними или наиболее часто встречающимися значениями соответствующего показателя.

2. Обрабатывать пропуски при подсчете расстояний. Пусть считается расстояние между x^i и x^j. Пусть x^i имеет пропущенное значение m-ого показателя. Тогда

$$x_m^i = f_{ij}(x^i, x^j, m),$$

   то есть функция учета может, вообще говоря, отличаться для разных объектов и разных пропущенных показателей.

Дальнейшая работа

Следующие шаги

1. Собрать данные.
2. Построить модель интегрального индикатора.
3. Выбрать показатель p расстояния Минковского и способ заполнения пропусков.
4. Написать код.