Расстояния, параметризованные размерами, и их аппроксимация

Выпускная квалификационная работа

Выполнила: Петренко Дарья Павловна, студентка 417 группы
Научный руководитель: к.ф.-м.н. Майсурадзе Арчил Ивериевич

МГУ им. Ломоносова

Москва, 2020

Понятие расстояния, параметризованного размером

X — произвольное множество («объекты»), S — частично упорядоченное множество с операцией сложения («размеры»)

Определение 1

Расстояние, параметризованное размером, — функция $\rho : X \times X \times S \to \mathbb{R}$, *удовлетворяющая системе аксиом:*

- *аксиомы расстояния* $\forall x_1, x_2 \in X$, $\forall s \in S$

D1. $\rho(x_1, x_1, s) = 0$ *(рефлексивность)*

D2. $\rho(x_1, x_2, s) = \rho(x_2, x_1, s)$ *(симметричность)*

D3. $\rho(x_1, x_2, s) \geqslant 0$ *(неотрицательность)*

- *аксиомы размера* $\forall x_1, x_2 \in X$, $\forall s_1, s_2 \in S$

S1. $s_1 \leqslant s_2 \Longrightarrow \rho(x_1, x_2, s_1) \leqslant \rho(x_1, x_2, s_2)$ *(монотонность)*

S2. $\rho(x_1, x_2, s_1 + s_2) \leqslant \rho(x_1, x_2, s_1) + \rho(x_1, x_2, s_2)$ *(неделимость)*

- Система аксиом (D1)—(D3), (S1)—(S2) непротиворечива.
- Система аксиом (D1)—(D3), (S1)—(S2) независима.

Усиления системы аксиом

Определение 2

Функция $\rho : X \times X \times S \to \mathbb{R}$ *называется псевдометрикой, параметризованной размером, если она удовлетворяет определению расстояния, параметризованного размером, и справедливо* $\forall x_1, x_2, x_3 \in X$, $\forall s \in S$

D4. $\rho(x_1, x_2, s) \leqslant \rho(x_1, x_3, s) + \rho(x_2, x_3, s)$ *(неравенство треугольника)*

Определение 3

Функция $\rho : X \times X \times S \to \mathbb{R}$ *называется метрикой, параметризованной размером, если она удовлетворяет определению псевдометрики, параметризованной размером, и справедливо* $\forall x_1, x_2 \in X$, $\forall s \in S$

D5. $\rho(x_1, x_2, s) = 0 \Rightarrow x_1 = x_2$ *(тождественность неразличимых)*

- Система аксиом (D1)—(D5), (S1)—(S2) непротиворечива.
- Аксиома (S1) независима от системы (D1)—(D5), (S2).
  Аксиома (S2) независима от системы (D1)—(D5), (S1).

Определение 4

Функция $\rho : X \times X \times S \to \mathbb{R}$ *называется* *различием, параметризованным размером*, *если* $\forall x_1, x_2 \in X$, $\forall s \in S$ *она удовлетворяет аксиомам*

D1. $\rho(x_1, x_1, s) = 0$ *(рефлексивность)*

D3. $\rho(x_1, x_2, s) \geqslant 0$ *(неотрицательность)*

Цели и задачи работы

Рассматриваются

- конечный набор объектов мощности N, причем объект характеризуется только своим индексом
- конечный набор размеров мощности K, для удобства дальнейших формулировок размеры — неотрицательные числа, пропорциональные соответствующим индексам в наборе

Дано: Δ — трехмерный тензор различий, параметризованных размером Элемент тензора δ_{ijk} содержит различие между i-тым и j-тым объектами, параметризованное k-тым размером

Цель работы: разработать методы аппроксимации (сжатия с потерями) тензора Δ, причем различия δ_{ijk}, параметризованные размерами, аппроксимируются расстояниями ρ_{ijk}, параметризованными размерами.

Оценка точности аппроксимации: взвешенное среднеквадратическое отклонение $Stress = \sum_i \sum_j \sum_k w_{ijk}(\delta_{ijk} - \rho_{ijk})^2$, где w_{ijk} — веса

Постановка задачи многомерного шкалирования

Многомерное шкалирование — это класс задач аппроксимации обычных различий обычными расстояниями

Рассматривается:

- конечный набор объектов мощности N, причем объект характеризуется только своим индексом

Дано: одна или несколько (двумерных) матриц $\Delta_k \in \mathbb{R}^{N \times N}, \ k = 1, \ldots, K$

$\delta_{ij,k}$ содержит различие между i-тым и j-тым объектами в k-той матрице

Задача: $\forall k$ найти представление объектов $x_1^k, \ldots, x_N^k$ в метрическом пространстве такое, чтобы $\text{dist}(x_i^k, x_j^k)$ оптимально согласовалось с $\delta_{ij,k}$

Постановка задачи многомерного шкалирования

Класс задач разбивается на подклассы:

- исходное пространство евклидово/ иное метрическое/ неметрическое
- пространство решений евклидово/ иное метрическое
- требуется согласование значений/ рангов
- требуется точное/ приближенное равенство
- одна/ несколько входных матриц различий

В частности,

- требуется согласование значений — метрическое многомерное шкалирование
- требуется согласование рангов — неметрическое многомерное шкалирование
- входных матриц несколько — многомерное шкалирование индивидуальных различий

Методика работы

Построим работу в соответствии с идеологией многомерного шкалирования.

- Постановка задачи условной оптимизации:

  1. Модель расстояния, параметризованного размером

  2. Достаточные условия выполнения аксиом для модели

  3. Функционал качества аппроксимации - всегда взвешенное среднеквадратическое отклонение

- Разработка метода оптимизации для поставленной задачи

Модель пропорциональных расстояний

- S - частично упорядоченное множество размеров с операцией сложения
- X — множество с расстоянием dist
- $r : S \to \mathbb{R}$

$$\rho(x_1, x_2, s) = r(s)\,\mathrm{dist}(x_1, x_2)$$

Теорема 1

Если dist — *расстояние, а* $r(s)$ *удовлетворяет*

$$\forall s \in S \quad r(s) \geqslant 0 \quad \textit{(неотрицательность)}$$

$$\forall s_1, s_2 \in S \quad s_1 \leqslant s_2 \Rightarrow r(s_1) \leqslant r(s_2) \quad \textit{(монотонность)}$$

$$\forall s_1, s_2 \in S \quad r(s_1 + s_2) \leqslant r(s_1) + r(s_2) \quad \textit{(субаддитивность)}$$

то функция ρ — *расстояние, параметризованное размером.*

Модель пропорциональных конфигураций

- S - частично упорядоченное множество размеров с операцией сложения
- X - множество с расстоянием dist
- для $\forall x \in X$ определена операция умножения на скаляр
- $r : S \to \mathbb{R}$

$$\rho(x_1, x_2, s) = \operatorname{dist}\left(r(s)x_1, r(s)x_2\right)$$

Предложение 1

Для случая $\operatorname{dist} = \operatorname{eucl}$ *модель пропорциональных конфигураций эквивалентна модели пропорциональных расстояний.*

Метод 1

- модель пропорциональных расстояний с $\text{dist} = \text{eucl}$
- входное пространство произвольное
- $X = [x_1, \ldots, x_N]^T$ — общая точечная конфигурация
- $r \in \mathbb{R}^K$ — вектор весов
- при условиях теоремы 1 требуется минимизировать

$$S(r, X) = \sum_{k=1}^{s} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ijk} (r_k \,\text{eucl}(x_i, x_j) - \delta_{ijk})^2$$

- решение — условная оптимизация вариационной верхней оценки S(X,r)

Модель индивидуальных различий

- X — множество с расстоянием dist
- $x \in X$ — векторы размерности L
- $\forall x \in X$ определена операция умножения на матрицу $\in \mathbb{R}^{L \times L}$
- S - частично упорядоченное множество размеров с операцией сложения
- $r : S \to \mathbb{R}^L$

$$\rho(x_1, x_2, s) = \operatorname{dist}\Big(x_1 \times \operatorname{diag}(r(s)), x_2 \times \operatorname{diag}(r(s))\Big)$$

Теорема 2

Если в модели индивидуальных различий $\operatorname{dist} = \operatorname{eucl}$, *а* r *удовлетворяет* $\forall s_1, s_2 \in S$

$$r(s_1) \geqslant 0 \quad \textit{(неотрицательность)}$$

$$s_1 \leqslant s_2 \;\Rightarrow r(s_1) \leqslant r(s_2) \quad \textit{(монотонность)}$$

$$r(s_1 + s_2) \leqslant r(s_1) + r(s_2) \quad \textit{(субаддитивность)}$$

то ρ — *расстояние, параметризованного размером.*

Метод 2

- модель индивидуальных различий с dist $=$ eucl
- входное пространство произвольное
- $X = [x_1, \ldots, x_N]^T$ — общая точечная конфигурация
- $r = [r_1, \ldots, r_K]^T \in \mathbb{R}^{K \times L}$ — векторы весов
- при условиях теоремы 2 требуется минимизировать

$$S(X, r) = \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ijk} \Big(\mathrm{eucl}\left(x_i \times \mathrm{diag}(r_k), x_j \times \mathrm{diag}(r_k) \right) - \delta_{ijk} \Big)^2$$

- решение — условная оптимизация вариационной верхней оценки S(X,r)

Модель рядов

- X — произвольное множество с функцией различия α
- S — частично упорядоченное множество с операцией сложения
- $r : X \times S \to \mathbb{R}$

$$\rho(x_1, x_2, s) = \alpha(x_i, x_j) r(x_i, s) + \alpha(x_j, x_i) r(x_j, s)$$

Теорема 3

Если для α верно

$$\alpha(x_i, x_j) \geqslant 0 \quad \textit{(рефлексивность)}$$

$$\alpha(x_i, x_i) = 0 \quad \textit{(неотрицательность)}$$

а r удовлетворяет

$$r(x_i, s) \geqslant 0 \quad \textit{(неотрицательность)}$$

$$s_1 \leqslant s_2 \Rightarrow r(x_i, s_1) \leqslant r(x_i, s_2) \quad \textit{(монотонность)}$$

$$r(x_1, s_1 + s_2) \leqslant r(x_1, s_1) + r(x_1, s_2) \quad \textit{(субаддитивность)}$$

то ρ — расстояние, параметризованное размером.

Метод 3

- модель рядов
- входное пространство произвольное
- $\alpha \in \mathbb{R}^{N \times N}$ - матрица попарных различий между объектами
- $r \in \mathbb{R}^{N \times K}$
- при условиях теоремы 3 требуется минимизировать

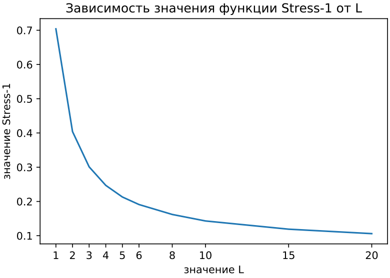
$$S(r, \alpha) = \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_{ij} r_{ik} + \alpha_{ji} r_{jk} - \delta_{ijk})^2$$

- решение — условная минимизация квадратичного функционала S(X)

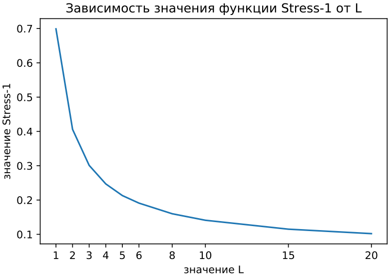
Результаты есперимeнтов

Эксперименты на реальных данных о задержках передачи сообщений между процессами многопроцессорной системы «Ломоносов», $\Delta \in \mathbb{R}^{78\times78\times100}$

Метод 1: $(NL + K)$ параметров



Метод 2: $(NL + KL)$ параметров



Метод 3: $(N^2 + NK)$ параметров (не зависит от L), $\quad$ Stress-1 $= 0.069$

Ошибку принято нормировать: Stress-1 $= \dfrac{\sum_i \sum_j \sum_k (\delta_{ijk} - \rho_{ijk})^2}{\sum_i \sum_j \sum_k \rho_{ijk}^2}$

Результаты эспериментов

Сравнение методов между собой:

- Высокое качество аппроксимации для моделей с малым числом параметров:
  - Метод 1 аппроксимирует не хуже, чем метод 2
  - Хорошая аппроксимация моделями с небольшим значением L
- Метод 3 дает наилучшую аппроксимацию, но содержит наибольшее число параметров
- Метод 3 позволяет получить для каждого объекта интерпретируемый вектор (ряд), описывающий изменение свойств этого объекта в зависимости от размера.

Сравнение с результатами безусловной оптимизации функции стресса:

- качество аппроксимации почти совпадает
- для результатов безусловной оптимизации достаточные условия теорем, а также аксиомы расстояния, параметризованного размером, не выполняются

Заключение

- Введено формальное понятие расстояния, параметризованного размером. Исследована система аксиом.
- Предложены модели таких расстояний с разным количеством параметров. Для всех моделей получены достаточные условия выполнения аксиом.
- На основе этих моделей и условий поставлены задачи и разработаны методы аппроксимации (сжатия) трехмерного тензора различий, параметризованных размером.
- Эксперименты на реальных данных показали хорошее качество для моделей с малым числом параметров (т. е. высокую степень сжатия) и сопоставимость потерь с безусловными постановками задач.