

# Комбинирование отношений порядка для восстановления предпочтения на наборе объектов

М. П. Кузнецов, В. В. Стрижов  
Московский физико-технический институт

Всероссийская конференция  
«Математические методы распознавания образов»  
Светлогорск, 24 сентября 2015.

# Задача восстановления предпочтения

## Цель исследования

Решение задачи восстановления отношения предпочтения на наборе объектов, заданных порядковым признаковым описанием.

## Методика

Предлагается подход на основе конусного представления предпочтений, заданных на наборе объектов.

## Задачи

1. Разработать метод построения суммы конусов предпочтений для восстановления отношения предпочтения.
2. Предложить метод восстановления предпочтения с использованием порождающего представления конусов.
3. Разработать метод снижения пространства параметров конусной модели.

# Постановка задачи восстановления предпочтений

## Дано

- ▶ Набор объектов  $x_1, \dots, x_m \in X$ .
- ▶ Набор предпочтений  $z_1, \dots, z_n$  на  $X$ :  $z_j(x_i, x_k) = \mathbb{I}[x_i \succeq_j x_k]$ ,  
 $z_j$  — отношение частичного или линейного порядка.
- ▶ Целевое отношение предпочтения  $z_0(x_i, x_k) = \mathbb{I}[x_i \succeq x_k]$ .

## Требуется

Построить агрегированное отношение предпочтения  $z_f$ , задаваемое отображением  $f(x_i) \in \mathbb{R}$ ,

- ▶ Удовлетворяющее условию монотонности по всем  $z_1, \dots, z_n$ ,

$$x_i \succeq_1 x_k, \dots, x_i \succeq_n x_k \quad \rightarrow \quad f_i \geq f_k,$$

- ▶ наилучшим образом приближающее целевое предпочтение  $z_0$ :

$$S(X, z_f, z_0) \rightarrow \min,$$

где  $S(X, z_f, z_0)$  — функция ошибки, описывающая различие между отношениями  $z_f$  и  $z_0$ .

# Предметная область

- ▶ Область социального выбора:  $X$  — множество кандидатов,  $z_1, \dots, z_n$  — избиратели.
- ▶ Задача комбинирования ранжирований (rank aggregation).
- ▶ Обучение ранжированию:  $z_0$  — оценки ассессоров поисковой системы.
- ▶ Порядковая классификация:  $z_0$  задается конечным множеством меток классов.

# Задача категоризации видов Красной книги РФ

Данные: экспертная анкета

| Вид                  | Численность | Площадь ареала | Генетическое разнообразие | Категория |
|----------------------|-------------|----------------|---------------------------|-----------|
| Зеленый осетр        | 2           | 2              | 0                         | 1         |
| Ладожский сиг        | 0           | 2              | 1                         | 2         |
| Длиннопёрая паляя    | 3           | 1              | 0                         | 3         |
| Полярный медведь     | 3           | 3              | 0                         | 4         |
| Канадский песочник   | 2           | 1              | 0                         | 3         |
| Азовская белуга      | 1           | 3              | 1                         | 1         |
| Водяной орех         | 3           | 3              | 2                         | 2         |
| Омфалина гудзонская  | 2           | 2              | 0                         | 3         |
| Сахалинский осетр    | 1           | 2              | 1                         | 1         |
| Гадюка Динника       | 3           | 3              | 2                         | 2         |
| Амурский тигр        | 2           | 2              | 1                         | 2         |
| Тропический лишайник | 2           | 1              | 1                         | 5         |

Описание признаков

| Признак                   | Шкала                        |
|---------------------------|------------------------------|
| Численность               | 3 — высокая                  |
|                           | 2 — низкая                   |
|                           | 1 — критически низкая        |
|                           | 0 — неизвестно               |
| Площадь ареала            | 3 — большая                  |
|                           | 2 — ограниченная             |
|                           | 1 — крайне ограниченная      |
|                           | 0 — неизвестно               |
| Генетическое разнообразие | 3 — высокое                  |
|                           | 2 — низкое                   |
|                           | 1 — неизвестно               |
| Категория                 | 5 — наименее угрожаемые      |
|                           | 4 — в уязвимости             |
|                           | 3 — под угрозой исчезновения |
|                           | 2 — в критическом состоянии  |
|                           | 1 — вымершие в дикой природе |

Попарное доминирование признаков

|                           | Численность | Площадь ареала | Генетическое разнообразие |
|---------------------------|-------------|----------------|---------------------------|
| Численность               | 1           | 1              | 1                         |
| Площадь ареала            | 0           | 1              | 0                         |
| Генетическое разнообразие | 0           | 0              | 1                         |

## Методы восстановления предпочтения

Для множества объектов  $X$  и отношения  $z_j$  определена матрица предпочтений  $\mathbf{Z}_j$ :  $\mathbf{Z}_j(i, k) = z_j(x_i, x_k)$ .

Методы, основанные на построении комбинации матриц  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ :

1. [Cohen et al., 1999]: линейная оценка матрицы

$$\text{предпочтений } \hat{\mathbf{Z}} = \sum_{j=1}^n w_j \mathbf{Z}_j,$$

восстановление линейного порядка  $\mathbf{f}$  по матрице  $\hat{\mathbf{Z}}$ .

2. [Liu et al., 2007]: построение взвешенной

$$\text{комбинации } f(x_i) = \sum_{j=1}^n w_j r_{ij},$$

$$r_{ij} = \#\{k \mid x_i \succeq_j x_k\} = \sum_{k=1}^m \mathbf{Z}(i, k).$$

3. [Volkovs et al., 2012]: построение признакового пространства на основе SVD-разложения  $\mathbf{Z}_j = \mathbf{U}_j \mathbf{\Sigma}_j \mathbf{V}_j^T$ .

# Конусное представление предпочтений

Дано

- ▶ Набор объектов  $x_1, \dots, x_m \in X$ .
- ▶ Набор предпочтений  $z_1, \dots, z_n$  на  $X$ :  $z_j(x_i, x_k) = \mathbb{I}[x_i \succeq_j x_k]$
- ▶ Целевое отношение предпочтения  $z_0(x_i, x_k) = \mathbb{I}[x_i \succeq x_k]$ .

Определение: конус предпочтений

$\mathcal{X}$  — конус предпочтений, задаваемый полиэдральным представлением с матрицей  $\mathbf{A}$  размера  $m^2 \times m$ :

$$\mathcal{X} = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{0}\},$$

где строка матрицы  $\mathbf{A}$  вида  $[0, \dots, 0, -1_i, 0, \dots, 0, 1_k, 0, \dots, 0]$  соответствует неравенству  $x_i \succeq x_k$ .

1.  $\mathcal{X}_1, \dots, \mathcal{X}_n$  — конусы, соответствующие предпочтениям  $z_1, \dots, z_n$ .
2.  $\mathcal{Y}_0$  — конус, соответствующий целевому предпочтению  $z_0$ .

# Построение суммы конусов предпочтений

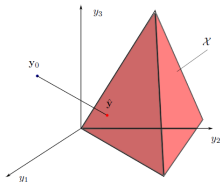
Конусная модель восстановления предпочтений

$$\mathbf{f} \in \mathcal{X}_f = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \dots \oplus \mathcal{X}_n, \quad \rho(\mathbf{z}_f, \mathbf{z}_0) = d(\mathcal{X}_f, \mathcal{Y}_0) \rightarrow \min.$$

Решение: проекция на допустимое множество значений

$$\hat{\mathbf{f}} = \arg \min_{\mathbf{f} \in \mathcal{X}_f, \mathbf{y}_0 \in \mathcal{Y}_0} \|\mathbf{f} - \mathbf{y}_0\|_2,$$

$$\hat{\mathbf{f}} = P_{\mathcal{X}_f}(\mathbf{y}_0).$$



Теорема (о представлении суммы конусов)

Суммой Минковского полиэдральных конусов  $\mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \dots \oplus \mathcal{X}_n$ , заданных матрицами  $\mathbf{A}_1, \dots, \mathbf{A}_n$ , является конус

$$\mathcal{X}_f = \{\mathbf{x} \mid \mathbf{A}^{(n)} \mathbf{x} \leq \mathbf{0}\},$$

задаваемый матрицей  $\mathbf{A}^{(n)} = \mathbf{V}_{n-1}^T \mathbf{A}^{(n-1)}$ , где  $\mathbf{V}_{n-1}$  — часть ФСР для уравнения с матрицей  $\begin{pmatrix} -\mathbf{A}^{(n-1)} \\ \mathbf{A}_n \end{pmatrix}$ .

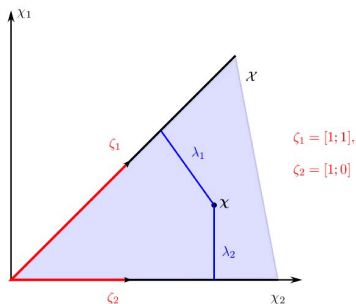


# Порождающее представление конуса

## Порождающее представление конуса

Полиэдральный конус  $\mathcal{X}$  допускает представление через конечный набор порождающих элементов  $\zeta_1, \dots, \zeta_k$ :

$$\mathcal{X} = \left\{ \sum_{k=1}^r \lambda_k \zeta_k \mid \lambda_k \geq 0 \right\}.$$



## Теорема (о порождающем представлении конуса)

Столбцы матрицы предпочтений  $\mathbf{Z}(i, k) = \mathbb{I}[x_i \succeq x_k]$  являются порождающими элементами конуса предпочтений,

$$\mathcal{X} \supset \{\mathbf{Z}\boldsymbol{\lambda} \mid \boldsymbol{\lambda} \in \mathbb{R}_+^m\}.$$

# Оценка параметров порождающего представления

Конусная модель восстановления предпочтений

$\mathbf{f} \in \mathcal{X}_f = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \dots \oplus \mathcal{X}_n$ ,  $\rho(z_f, z_0) = d(\mathcal{X}_f, \mathcal{Y}_0) \rightarrow \min$ .

Использование порождающего представления

Линейная конусная модель:  $\mathcal{X}_j = \{\mathbf{Z}_j \boldsymbol{\lambda}_j \mid \boldsymbol{\lambda}_j \in \mathbb{R}_+^m\}$ ,

$$\mathbf{f}(x_1, \dots, x_m) = \sum_{j=1}^n \mathbf{Z}_j \boldsymbol{\lambda}_j, \quad \boldsymbol{\lambda}_j \geq \mathbf{0}.$$

Минимизация расстояния между конусами:

$$(\hat{\boldsymbol{\lambda}}_1, \dots, \hat{\boldsymbol{\lambda}}_n) = \arg \min_{\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n \geq \mathbf{0}} \left\| \mathbf{Z}_0 \mathbf{1} - \sum_{j=1}^n \mathbf{Z}_j \boldsymbol{\lambda}_j \right\|_2.$$

Итеративный алгоритм оценки параметров

Шаг алгоритма — последовательное решение задач неотрицательной линейной регрессии

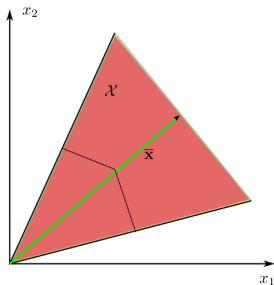
$$\hat{\boldsymbol{\lambda}}_j^t = \arg \min_{\boldsymbol{\lambda}_j \geq \mathbf{0}} \left\| \mathbf{Z}_0 \mathbf{1} - \sum_{j'=1}^{j-1} \mathbf{Z}_{j'} \hat{\boldsymbol{\lambda}}_{j'}^t - \sum_{j'=j+1}^m \mathbf{Z}_{j'} \hat{\boldsymbol{\lambda}}_{j'}^{t-1} - \mathbf{Z}_j \boldsymbol{\lambda}_j \right\|_2.$$

# Регуляризация конусной модели

Линейная конусная модель:

$$\mathbf{f}(X) = \sum_{j=1}^n \mathbf{z}_j \lambda_j, \quad \lambda_j \geq 0.$$

Рассмотрим в конусе  $\mathcal{X}$  центральную точку  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^n \mathbf{z}_j$ .



Теорема (о регуляризации конусной модели)

В случае замены каждого конуса  $\mathcal{X}_k = \{\sum \lambda_{jk} \mathbf{z}_{jk} \mid \lambda_k \geq \mathbf{0}\}$  его центральной точкой конусная модель представима в виде

$$\mathbf{f}(\mathbf{x}_1, \dots, \mathbf{x}_m) = \hat{\mathbf{Z}} \boldsymbol{\lambda}, \quad \hat{\mathbf{Z}} = \sum_{j=1}^n w_j \mathbf{z}_j,$$

при ограничениях  $w_j \geq 0$ ,  $\sum_{k=1}^m \lambda_k = 1$ ,  $\boldsymbol{\lambda} \geq \mathbf{0}$ .

# Оценка параметров регуляризованной модели

## Минимизация расстояния между конусами

Для регуляризованной модели задача минимизация расстояния между конусами  $\rho(\mathcal{X}_f, \mathcal{Y}_0)$  сводится к минимизации нормы разности матриц  $\hat{\mathbf{Z}}, \mathbf{Z}_0$ :

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \|\hat{\mathbf{Z}} - \mathbf{Z}_0\|_F^2 \propto -\tau(z_f, z_0).$$

## Алгоритм восстановления предпочтения

Алгоритм основывается на построении взвешенного графа предпочтений, описываемого матрицей смежности  $\hat{\mathbf{Z}}$ :

1. Оценка весов  $w_j$  в модели  $\hat{\mathbf{Z}} = \sum_{j=1}^n w_j \mathbf{Z}_j$ .
2. Оценка параметров  $\lambda$  и построение оценок объектов  $\mathbf{f}(x_1, \dots, x_n) = \hat{\mathbf{Z}}\lambda$ .

# Задача категоризации видов Красной книги РФ

Данные: экспертная анкета

| Вид                  | Численность | Площадь ареала | Генетическое разнообразие | Категория |
|----------------------|-------------|----------------|---------------------------|-----------|
| Зеленый осетр        | 2           | 2              | 0                         | 1         |
| Ладожский сиг        | 0           | 2              | 1                         | 2         |
| Длиннопёрая паляя    | 3           | 1              | 0                         | 3         |
| Полярный медведь     | 3           | 3              | 0                         | 4         |
| Канадский песочник   | 2           | 1              | 0                         | 3         |
| Азовская белуга      | 1           | 3              | 1                         | 1         |
| Водяной орех         | 3           | 3              | 2                         | 2         |
| Омфалина гудзонская  | 2           | 2              | 0                         | 3         |
| Сахалинский осетр    | 1           | 2              | 1                         | 1         |
| Гадюка Динника       | 3           | 3              | 2                         | 2         |
| Амурский тигр        | 2           | 2              | 1                         | 2         |
| Тропический лишайник | 2           | 1              | 1                         | 5         |

Описание признаков

| Признак                   | Шкала                        |
|---------------------------|------------------------------|
| Численность               | 3 — высокая                  |
|                           | 2 — низкая                   |
|                           | 1 — критически низкая        |
|                           | 0 — неизвестно               |
| Площадь ареала            | 3 — большая                  |
|                           | 2 — ограниченная             |
|                           | 1 — крайне ограниченная      |
|                           | 0 — неизвестно               |
| Генетическое разнообразие | 3 — высокое                  |
|                           | 2 — низкое                   |
|                           | 1 — неизвестно               |
| Категория                 | 5 — наименее угрожаемые      |
|                           | 4 — в уязвимости             |
|                           | 3 — под угрозой исчезновения |
|                           | 2 — в критическом состоянии  |
|                           | 1 — вымершие в дикой природе |

Попарное доминирование признаков

|                           | Численность | Площадь ареала | Генетическое разнообразие |
|---------------------------|-------------|----------------|---------------------------|
| Численность               | 1           | 1              | 1                         |
| Площадь ареала            | 0           | 1              | 0                         |
| Генетическое разнообразие | 0           | 0              | 1                         |

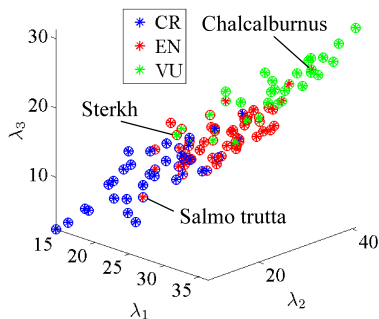
# Результаты категоризации

Ошибка — средняя потеря Хэмминга

$$L_H(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i|_H.$$

Категоризация Красной книги: сравнение алгоритмов.  
OW — регуляризованная конусная модель.

| Алгоритм | $L_H$        |
|----------|--------------|
| OW       | <b>0.52*</b> |
| Копулы   | 0.59         |
| CR       | 0.71         |
| Trees    | 0.55         |
| SVM      | 0.66         |
| kNN      | 0.72         |



# Порядковая классификация, данные UCI

Функция ошибки:

1. Средняя абсолютная ошибка,  $L_a(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{m} \sum_{i=1}^m [y_i \neq \hat{y}_i]$ ,
2. Средняя ошибка Хэмминга,  $L_H(\mathbf{y}, \hat{\mathbf{y}})$ .

Результаты на данных UCI: линейные признаки

| Данные   | Средняя абсолютная ошибка ( $\pm 0.01$ ) |      |       |              |      | Средняя ошибка Хэмминга ( $\pm 0.01$ ) |      |              |              |      |
|----------|--|------|-------|--------------|------|--|------|--------------|--------------|------|
|          | SVM                                      | POF  | Trees | OW           | KNN  | SVM                                    | POF  | Trees        | OW           | KNN  |
| Pyr      | <b>0.50*</b>                             | 0.62 | 0.61  | 0.54         | 0.55 | <b>0.64*</b>                           | 0.90 | 0.84         | 0.75         | 0.75 |
| CPU      | 0.44                                     | 0.44 | 0.47  | <b>0.42*</b> | 0.51 | 0.53                                   | 0.53 | 0.53         | <b>0.49*</b> | 0.61 |
| Boston   | <b>0.38*</b>                             | 0.48 | 0.41  | <b>0.39*</b> | 0.47 | <b>0.46*</b>                           | 0.65 | <b>0.47*</b> | <b>0.46*</b> | 0.62 |
| Computer | <b>0.32*</b>                             | 0.71 | 0.38  | 0.34         | 0.60 | <b>0.35*</b>                           | 1.36 | 0.41         | 0.39         | 0.90 |
| Abalone  | <b>0.53*</b>                             | 0.59 | 0.57  | 0.56         | 0.60 | <b>0.78*</b>                           | 0.92 | <b>0.77*</b> | 0.81         | 0.88 |

Результаты на данных UCI: порядковые признаки

| Данные   | Средняя абсолютная ошибка ( $\pm 0.01$ ) |              |              |              |              | Средняя ошибка Хэмминга ( $\pm 0.01$ ) |              |              |              |              |
|----------|--|--------------|--------------|--------------|--------------|--|--------------|--------------|--------------|--------------|
|          | SVM                                      | POF          | Trees        | OW           | KNN          | SVM                                    | POF          | Trees        | OW           | KNN          |
| Pyr      | 0.57                                     | 0.58         | 0.60         | 0.62         | <b>0.49*</b> | <b>0.71*</b>                           | 0.77         | 0.79         | 0.79         | 0.76         |
| CPU      | 0.51                                     | <b>0.39*</b> | 0.47         | <b>0.40*</b> | 0.43         | 0.65                                   | <b>0.45*</b> | 0.56         | 0.47         | 0.51         |
| Boston   | <b>0.40*</b>                             | 0.48         | <b>0.40*</b> | 0.43         | <b>0.41*</b> | 0.49                                   | 0.68         | <b>0.46*</b> | 0.50         | 0.51         |
| Computer | 0.44                                     | 0.69         | 0.41         | <b>0.37*</b> | 0.45         | 0.53                                   | 1.38         | <b>0.45*</b> | <b>0.44*</b> | 0.55         |
| Abalone  | 0.78                                     | <b>0.59*</b> | <b>0.57*</b> | <b>0.58*</b> | <b>0.59*</b> | 1.78                                   | 0.92         | <b>0.76*</b> | 0.85         | 0.89         |
| Cars     | 0.19                                     | 0.19         | 0.08         | 0.16         | <b>0.06*</b> | 0.24                                   | 0.26         | <b>0.08*</b> | 0.19         | <b>0.07*</b> |
| RedBook  | 0.56                                     | 0.61         | <b>0.50*</b> | <b>0.49*</b> | 0.59         | 0.66                                   | 0.74         | <b>0.55*</b> | 0.59         | 0.72         |

## Основные результаты

1. Предложен метод построения суммы конусов предпочтений для решения задачи восстановления отношения предпочтения на множестве объектов, заданных порядковым признаковым описанием.
2. Предложен метод восстановления предпочтения с использованием порождающего представления конусов.
3. Предложен метод снижения размерности пространства параметров конусной модели.
4. Решена прикладная задача категоризации таксонов Красной книги по экспертным оценкам, предоставленным министерством природных ресурсов РФ.

## Публикации по теме работы

1. M.P. Kuznetsov, V.V. Strijov. Methods of expert estimations concordance for integral quality estimation // Expert Systems with Applications, 41(4):1988-1996, 2014.
2. М. П. Кузнецов, В. В. Стрижов, М.М Медведникова. Алгоритм многоклассовой классификации объектов, описанных в ранговых шкалах. // Научно-технический вестник СПб ГПУ. Информатика. Телекоммуникации. Управление, 5, 2012.