Построение интегральных индикаторов в задачах с порядковыми признаками

М. П. Кузнецов

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

Научный руководитель к.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

Москва, 2013 г.

Цели и методы

Цели работы

- Предложить алгоритм построения интегральных индикаторов объектов, описанных в смешанных (линейных, порядковых, номинальных) шкалах.
- При построении интегральных индикаторов необходимо учитывать
  - Экспертные предпочтения на множестве объектов,
  - экспертные предпочтения на множестве признаков, либо собственную размерность пространства описаний.

Базовые методы

- В. В. Стрижов. Уточнение экспертных оценок, выставленных в ранговых шкалах, с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов, 2011, 77(7) — 72-78.
- J. Furkranz, E. Hullermeier. Preference learning. Springer,2011.

Определения и обозначения

Задана выборка $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$.

- $\mathbf{X} = [\chi_1, \dots, \chi_n]$ — матрица описаний объектов,

$$\text{признаки } \chi_j \subset \mathbb{R}^m, \quad j \in \mathcal{J} = \{1, \dots, n\}.$$

- $\mathbf{y} \subset \mathbb{R}^m$ — множество интегральных индикаторов.
- Линейная шкала: $\quad \chi_j \in \mathbb{R}^m, \quad j \in \mathcal{J}.$
- Порядковая шкала (линейный порядок):

$$\chi_j = \{\mathbf{x}_j \in \mathbb{R}^m_+ | \quad x_{j1} \geqslant x_{j2} \geqslant \dots \geqslant x_{jm}\}.$$

- Порядковая шкала (частичный порядок):

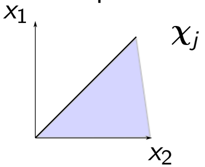
$$\chi_j = \{\mathbf{x}_j \in \mathbb{R}^m_+ | \quad x_{jk_1} \geqslant x_{jk_2}, \dots\}.$$

Порядковая шкала

- Признак χ_j, заданный в порядковой шкале,

$$\chi_j = \{\mathbf{x}_j \in \mathbb{R}^m_+ | \quad x_{j1} \geqslant x_{j2} \geqslant \ldots \geqslant x_{jm}\},$$

является выпуклым многогранным конусом в $\mathbb{R}^m$.



- Конус χ_j соответствует частичному порядку, заданному на множестве значений j-го признака.

Задача построения интегральных индикаторов

- Модель
$$\mathbf{y}_1 = f(\mathbf{X}, \mathbf{w}), \quad \mathbf{y}_1 \subset \mathbb{R}^m,$$
оптимальные параметры $\hat{w}$ минимизируют ошибку
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} S(f(\mathbf{X}, \mathbf{w}), \mathbf{y}_0).$$
где $\mathbf{y}_0 \subset \mathbb{R}^m$ — экспертная оценка множества интегральных индикаторов.
- Предлагается параметризовать множества $\chi_j \subset \mathbb{R}^m$, а также множество $\mathbf{y}_0 \subset \mathbb{R}^m$.

Разложение точки конуса по образующим конуса

Теорема: для любой точки $\mathbf{x}$ конуса $\boldsymbol{\chi}$ справедливо разложение

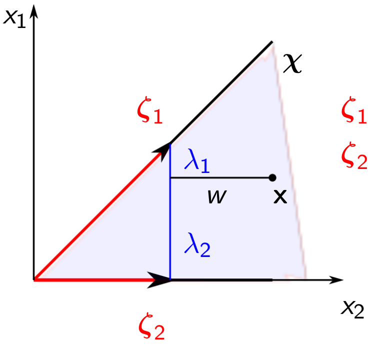
$$\mathbf{x} = w \sum_{k=1}^{L} \lambda_k \boldsymbol{\zeta}_k, \quad w \geqslant 0, \quad \sum_{k=1}^{L} \lambda_k = 1, \quad \lambda_k \geq 0,$$

где L — мощность множества значений признака $\boldsymbol{\chi}$,
$\boldsymbol{\zeta}_k$ — образующая конуса, соответствующая признаку $\boldsymbol{\chi}$:

$$\zeta_k(i) = \begin{cases} 1, & \text{если } x_i \geqslant x_k, \\ 0, & \text{иначе.} \end{cases}$$

Иллюстрация разложения точки по образующим конуса

$$\mathbf{x} = w \sum_{k=1}^{L} \lambda_k \zeta_k, \quad w \geqslant 0, \quad \sum_{k=1}^{L} \lambda_k = 1, \quad \lambda_k \geq 0.$$



$\zeta_1 = [1; 1],$

$\zeta_2 = [1; 0]$

Иллюстрация бинаризации порядковой шкалы

Рассмотрим пример частично упорядоченного множества

$$\chi = \{x_1, x_2, x_3 | \quad x_1 \geqslant x_2, x_1 \geqslant x_3\}.$$

$$\begin{tikzcd} x_1 \arrow[r] \arrow[dr] & x_2 \\ & x_3 \end{tikzcd}$$

Матрица $\boldsymbol{Z} = [\zeta_1, \zeta_2, \zeta_3]$, соответствующая этому графу:

$$\boldsymbol{Z} = \begin{array}{c} \begin{array}{ccc} \zeta_1 & \zeta_2 & \zeta_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array},$$

$$\chi \ni \mathbf{x} = \boldsymbol{\lambda} \boldsymbol{Z}, \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \quad \|\boldsymbol{\lambda}\| = 1.$$

Модель построения интегральных индикаторов

- Модель построения интегральных индикаторов:

$$\mathbf{y}_1 = f([\chi_1, ..., \chi_n], \mathbf{w}).$$

- Обобщение линейной модели на случай порядковых признаков:

$$\mathbf{y}_1 = w_1 \boldsymbol{\lambda}_1 \boldsymbol{Z}_1 + ... + w_n \boldsymbol{\lambda}_n \boldsymbol{Z}_n,$$

$$w_j \in \mathbb{R}, \qquad \boldsymbol{\lambda}_j \in \Lambda = \{\boldsymbol{\lambda} | \boldsymbol{\lambda} \geqslant \mathbf{0}, \|\boldsymbol{\lambda}\|_1 = 1\}.$$

Оценка параметров модели

- Задача поиска оптимальных параметров модели, доставляющих минимум функции ошибки

$$S(\mathbf{f}(\mathbf{X}, \mathbf{w}), \mathbf{y}_0) \to \min_{\mathbf{w} \in \mathbb{R}^{\boldsymbol{n}}},$$

обобщается на случай порядковых признаков:

$$S(\mathbf{f}([\chi_1, \dots, \chi_n], \mathbf{w}), \mathbf{y}_0) \to \min_{\substack{\mathbf{w} \in \mathbb{R}^{\boldsymbol{n}};\\ \boldsymbol{\lambda_1}, \dots, \boldsymbol{\lambda_n}, \boldsymbol{\lambda_0} \in \Lambda}}.$$

- В линейном случае:

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}}_1, \dots, \hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\lambda}}_0) = \mathop{\arg\min}_{\substack{\mathbf{w} \in \mathbb{R}^{\boldsymbol{n}};\\ \boldsymbol{\lambda_1}, \dots, \boldsymbol{\lambda_n}, \boldsymbol{\lambda_0} \in \Lambda}} \|\boldsymbol{\lambda}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} w_j \boldsymbol{\lambda}_j \boldsymbol{Z}_j\|.$$

Итеративный алгоритм минимизации параметров

Идея: итеративно вычислять $\mathbf{w}$ при фиксированных $\hat{\boldsymbol{\lambda}}_1, \dots, \hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\lambda}}_0$, затем вычислять $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n, \boldsymbol{\lambda}_0$ при фиксированном $\hat{\mathbf{w}}$.

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}}_1, \dots, \hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\lambda}}_0) = \mathop{\arg\min}_{\substack{\mathbf{w} \in \mathbb{R}^n; \\ \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n, \boldsymbol{\lambda}_0 \in \Lambda}} \|\boldsymbol{\lambda}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} w_j \boldsymbol{\lambda}_j \boldsymbol{Z}_j\|.$$

- Начальное приближение:

$$\hat{\boldsymbol{\lambda}}_{jk} = 1/L_j, \qquad j \in \mathcal{J}, \quad k = 1, \dots, L_j.$$

- Шаг $t+1$:
  1. $\hat{\mathbf{w}} := \mathop{\arg\min}_{\mathbf{w} \in \mathbb{R}} \|\hat{\boldsymbol{\lambda}}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} w_j \hat{\boldsymbol{\lambda}}_j \boldsymbol{Z}_j\|,$
  2. $\hat{\boldsymbol{\lambda}}_j = \mathop{\arg\min}_{\boldsymbol{\lambda}_j \in \Lambda} \|\hat{\boldsymbol{\lambda}}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} \hat{w}_j \hat{\boldsymbol{\lambda}}_j \boldsymbol{Z}_j\|$, для всех $j \in \mathcal{J}$.

Теорема о сходимости

- Количество свободных параметров в модели с порядковыми признаками:

$$\Theta = L_0 - 1 + L_1 - 1 + \ldots + L_n - 1 + n = \sum_{j=0}^{n} L_j - 1.$$

- Теорема: в случае обобщения линейной модели,

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}}_1, \ldots, \hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\lambda}}_0) = \underset{\substack{\mathbf{w} \in \mathbb{R}^n; \\ \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n, \boldsymbol{\lambda}_0 \in \Lambda}}{\arg\min} \|\boldsymbol{\lambda}_0 \mathbf{Z}_0 - \sum_{j=1}^{n} w_j \boldsymbol{\lambda}_j \mathbf{Z}_j\|,$$

алгоритм находит оптимальное решение не более чем за Θ итераций.

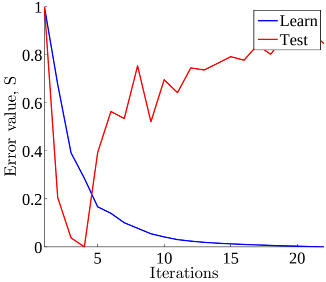
Категоризация редких видов красной книги РФ

- Категории редких видов:
  1 — в критическом состоянии, 2 — под угрозой исчезновения, 3 — в уязвимости.
- Примеры признаков в анкете:
  РП — Размер Популяции, ПА — Площадь Ареала, ГР — Генетическое разнообразие.

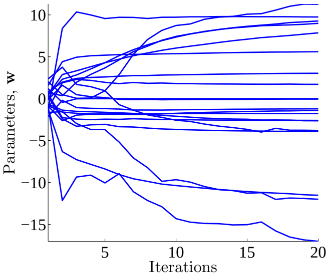
Вид	РП	ПА	ГР	y_0	y_1
Зеленый осетр	2	2	0	1	?
Ладожский сиг	2	2	1	2	?
Длиннопёрая палия Световидова	3	1	0	3	?
Полярный медведь	3	3	0	2	?
Канадский песочник	2	1	0	3	?
Шизофрагма гортензиевидная	1	1	1	1	?
Тропический лишай	2	1	1	3	?

Сходимость параметров

- Размер обучающей выборки: $m = 70$.
- Количество параметров: $\Theta = 40$.



(a) Значение ошибки



(b) Параметры модели, **w**

Восстановленная категоризация

РП — Размер Популяции, ПА — Площадь Ареала, ГР — Генетическое разнообразие.

Вид	РП	ПА	ГР	y_0	y_1
Зеленый осетр	2	2	0	1	2
Ладожский сиг	2	2	1	2	2
Длиннопёрая палия Световидова	3	1	0	3	3
Полярный медведь	3	3	0	2	1
Канадский песочник	2	1	0	3	2
Шизофрагма гортензиевидная	1	1	1	1	1
Тропический лишай	2	1	1	3	2

Результаты

- Предложен итеративный алгоритм построения интегральных индикаторов объектов, описанных в смешанных (линейных, порядковых) шкалах.
- Доказана теорема о максимальном количестве итераций алгоритма для нахождения оптимального решения.
- Установлено соответствие между конусом, описывающим частично упорядоченное множество, и матрицей инцидентности графа, соответствующего этому множеству.

Публикации ВАК:

1. Кузнецов М.П., Стрижов В.В., Медведникова М.М. Алгоритм многоклассовой классификации объектов, описанных в ранговых шкалах // НТВ СпбГПУб 2012.
2. Стрижов В.В., Кузнецов М.П., Рудаков К.В. Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах // Математическая биология и биоинформатика, 2012.
3. Кузнецов М.П. Уточнение ранговых экспертных оценок с использованием монотонной интерполяции и конусов // Труды МФТИ, 2013.
4. Медведникова М.М., Стрижов В.В., Кузнецов М.П. Алгоритм многоклассовой монотонной Парето-классификации с выбором признаков // Известия Тульского государственного университета, Естественные науки, 2012