Построение моделей обучения по предпочтениям с использованием порядковых экспертных оценок

М. П. Кузнецов

Диссертация на соискание ученой степени
кандидата физико-математических наук
05.13.18 — Математическое моделирование, численные методы и комплексы программ
научный руководитель д.ф.-м.н. В.В. Стрижов

Московский физико-технический институт
32 августа 2016 г.

Задача обучения по предпочтениям

Цели исследования

1. Разработка новых подходов к моделированию объектов, заданных порядковым описанием.
2. Разработка и обоснование эффективных вычислительных методов распознавания объектов, заданных линейным и порядковым описанием.

Задачи

1. Разработать подход к моделированию порядковых признаков с использованием полиэдрального представления порядковых данных.
2. Разработать алгоритм построения модели предпочтений на множестве объектов, заданных порядковым признаковым описанием.
3. Разработать метод решения задачи порядковой классификации объектов, заданных линейным и порядковым описанием.
4. Разработать методы моделирования и согласования линейных и порядковых экспертных оценок.
5. Реализовать алгоритмы распознавания порядковых объектов и провести вычислительный эксперимент для установления границ применимости методов.

Постановка задачи восстановления предпочтений

Дано

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

Требуется

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

- удовлетворяющее условию монотонности по всем $z_1, \ldots, z_n$,
$$x_i \geq_1 x_k, \ldots, x_i \geq_n x_k \quad \rightarrow \quad f_i \geq f_k,$$
- 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, $z_1, \ldots, z_n$ — избиратели.
- Задача комбинирования ранжирований: X — множество документов, $z_1, \ldots, z_n$ — ответы поисковых систем.
- Обучение ранжированию: z_0 — оценки асессоров поисковой системы.
- Порядковая классификация: X — множество объектов, целевое отношение 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 палия	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 $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$:

1. [Cohen et al., 1999]: линейная оценка матрицы предпочтений $\hat{\mathbf{Z}} = \sum\limits_{j=1}^{n} w_j \mathbf{Z}_j$, восстановление линейного порядка $\mathbf{f}$ по матрице $\hat{\mathbf{Z}}$.
2. [Liu et al., 2007]: построение взвешенной комбинации $f(x_i) = \sum\limits_{j=1}^{n} w_j r_{ij}$, $r_{ij} = \#\{k \mid x_i \succeq_j x_k\} = \sum\limits_{k=1}^{m} \mathbf{Z}(i,k)$.
3. [Volkovs et al., 2012]: построение признакового пространства на основе SVD-разложения $\mathbf{Z}_j = \mathbf{U}_j \boldsymbol{\Sigma}_j \mathbf{V}_j^{\mathsf{T}}$.

Конусное представление предпочтений

Дано

- Набор объектов $x_1, \ldots, x_m \in X$.
- Набор предпочтений $z_1, \ldots, z_n$ 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 $[0, \ldots, 0, -1_i, 0, \ldots, 0, 1_k, 0, \ldots, 0]$ соответствует неравенству $x_i \succeq x_k$.

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

Построение суммы конусов предпочтений

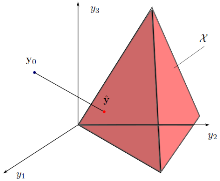
Конусная модель восстановления предпочтений

$$\mathbf{f} \in \mathcal{X}_f = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \ldots \oplus \mathcal{X}_n, \quad S(X, z_f, z_0) = d(\mathcal{X}_f, \mathcal{Y}_0) \to \min .$$

Решение: проекция на допустимое множество значений

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

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



Алгоритм построения суммы конусов

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

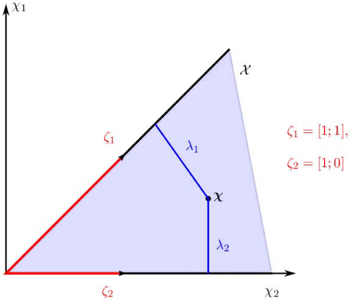
$$\mathcal{X}_f = \{\mathbf{x} \mid \mathbf{A}^{(n)}\mathbf{x} \leqslant \mathbf{0}\},$$

задаваемый матрицей $\mathbf{A}^{(n)} = \mathbf{V}_{n-1}^{\mathsf{T}}\mathbf{A}^{(n-1)}$, 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

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

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



Теорема (о порождающем представлении конуса), [Кузнецов: 2013]

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

$$\mathbf{f} \in \mathcal{X}_f = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \ldots \oplus \mathcal{X}_n, \quad S(X, z_f, z_0) = d(\mathcal{X}_f, \mathcal{Y}_0) \to \min.$$

Использование порождающего представления

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

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

Минимизация расстояния между конусами:

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

Итеративный алгоритм оценки параметров

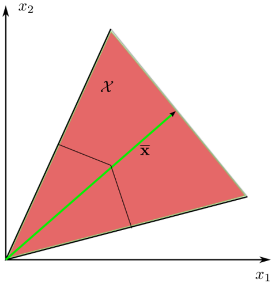
Шаг алгоритма — последовательное решение задач неотрицательной линейной регрессии

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

Регуляризация конусной модели

Линейная конусная модель:

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



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), [Кузнецов: 2014]

В случае замены каждого конуса $\mathcal{X}_k = \{\sum \lambda_{jk}\mathbf{z}_{jk} \mid \boldsymbol{\lambda}_k \geq \mathbf{0}\}$ 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 $w_j \geq 0, \; \sum^{m} \lambda_k = 1, \; \boldsymbol{\lambda} \geq \mathbf{0}.$

Оценка параметров регуляризованной модели

Следствие: 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 $\hat{\mathbf{Z}} = \sum\limits_{j=1}^{n} w_j \mathbf{Z}_j$.
2. Оценка параметров $\boldsymbol{\lambda}$ и построение оценок объектов $\mathbf{f}(x_1, ..., x_n) = \hat{\mathbf{Z}}\boldsymbol{\lambda}$.

Согласование экспертных оценок

Дано

- Набор объектов $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$.
- Матрица плана $\mathbf{X}$ размера $m \times n$, $m \geq n$, полного ранга.
- Линейная модель построения оценок объектов $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$.
- Экспертные оценки целевой переменной и параметров $\mathbf{w}$.
  1. Линейные экспертные оценки: $\mathbf{y}_0 \in \mathbb{R}^m$, $\mathbf{w}_0 \in \mathbb{R}^n$.
  2. Порядковые экспертные оценки: $\mathcal{Y}_0 \subset \mathbb{R}^m$, $\mathcal{W}_0 \subset \mathbb{R}^n$.

Определение согласованности оценок

Оценки $\mathbf{y}$ и $\mathbf{w}$ называются согласованными, если для них выполняются следующие условия:

$$\mathbf{y} = \mathbf{X}\mathbf{w}, \quad \mathbf{w} = \mathbf{X}^+\mathbf{y}.$$

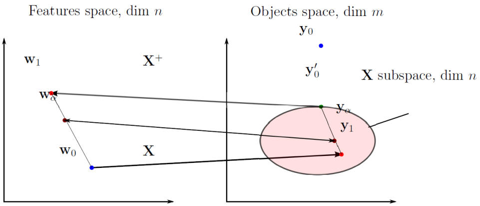
Задача согласования оценок

Требуется получить согласованную пару оценок $\hat{\mathbf{y}}$, $\hat{\mathbf{w}}$, ближайшую к паре экспертных оценок.

Согласование линейных экспертных оценок

Линейные экспертные оценки: $\mathbf{y}_0 \in \mathbb{R}^m$, $\mathbf{w}_0 \in \mathbb{R}^n$, $\mathbf{y}_0' = \mathbf{X}\mathbf{X}^+\mathbf{y}_0$.

Методы получения согласованных оценок



- Метод α-согласования:

$$\mathbf{w}_\alpha = \alpha\mathbf{w}_0 + (1-\alpha)\mathbf{X}^+\mathbf{y}_0', \quad \mathbf{y}_\alpha = (1-\alpha)\mathbf{y}_0' + \alpha\mathbf{X}\mathbf{w}_0.$$

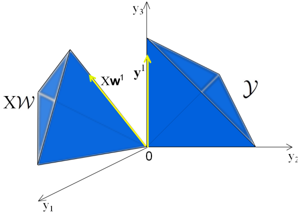
- Метод γ-согласования:

$$\mathbf{w}_\gamma = (\mathbf{X}^\mathsf{T}\mathbf{X} + \gamma^2 I_n)^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y}_0' + \gamma^2\mathbf{w}_0), \quad \mathbf{y}_\gamma = \mathbf{X}\mathbf{w}_\gamma.$$

Согласование порядковых экспертных оценок

Порядковые экспертные оценки: $\mathcal{Y}_0 \subset \mathbb{R}^m$, $\mathcal{W}_0 \subset \mathbb{R}^n$.

Задача: поиск ближайших векторов в конусах $\mathcal{Y}_0$ и $\mathcal{W}_0$



$$(\hat{\mathbf{w}}, \mathbf{y}_1) = \underset{\mathbf{w}\in\mathcal{W}_0, \mathbf{y}\in\mathcal{Y}_0}{\arg\min} \|\mathbf{X}^+\mathbf{y} - \mathbf{w}\|_2,$$

$$\|\mathbf{X}^+\mathbf{y}\|_2 = 1, \; \|\mathbf{w}\|_2 = 1.$$

Методы получения согласованных оценок, [Кузнецов: 2013]

- Полиэдральное представление конусов:

Задача $2k$:		Задача $2k+1$:	
minimize	$\|\mathbf{X}^+\mathbf{a} - \mathbf{w}^{(2k)}\|_2$	minimize	$\|\mathbf{X}^+\mathbf{y}^{(2k+1)} - \mathbf{b}\|_2$
subject to	$(\mathbf{X}^+\mathbf{a})^\mathsf{T}\mathbf{X}^+\mathbf{a} = 1,$	subject to	$\mathbf{b}^\mathsf{T}\mathbf{b} = 1,$
	$\mathbf{A}^m\mathbf{a} \leqslant 0.$		$\mathbf{A}^n\mathbf{b} \leqslant 0.$

- Порождающее представление конусов:

Задача $2k$:	Задача $2k+1$:
$\hat{\boldsymbol{\lambda}}_0 = \min\limits_{\boldsymbol{\lambda}_0 \geq \mathbf{0},\ \|\boldsymbol{\lambda}_0\|=1} \|\mathbf{Z}_0\boldsymbol{\lambda}_0 - \mathbf{X}_w\hat{\boldsymbol{\lambda}}_w\|_2^2$	$\hat{\boldsymbol{\lambda}}_w = \min\limits_{\boldsymbol{\lambda}_w \geq \mathbf{0},\ \|\boldsymbol{\lambda}_w\|=1} \|\mathbf{Z}_0\hat{\boldsymbol{\lambda}}_0 - \mathbf{X}_w\boldsymbol{\lambda}_w\|_2^2$

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 палия	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

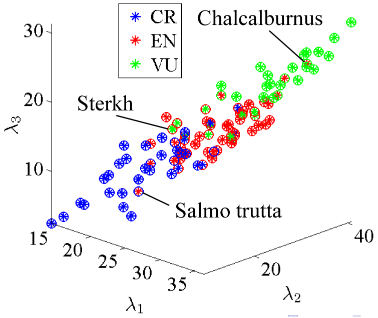
Результаты категоризации

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	**0.52** *
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

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

Результаты на данных UCI: порядковые признаки

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

Результаты, выносимые на защиту

1. Предложен метод моделирования объектов, заданных порядковым описанием, с использованием полиэдрального представления предпочтений на множестве объектов.
2. Разработан метод построения модели предпочтений объектов на основе суммы конусов предпочтений.
3. Предложен численный метод решения задачи порядковой классификации объектов на основе оценивания матрицы попарного доминирования объектов.
4. Разработаны методы моделирования и согласования порядковых экспертных оценок объектов, заданных линейным признаковым описанием.
5. Разработан программный комплекс, включающий в себя методы распознавания объектов, заданных линейным и порядковым описанием. Проведены вычислительные эксперименты, подтверждающие адекватность методов.

Список работ автора по теме диссертации

Публикации ВАК

1. M.M. Stenina, M.P. Kuznetsov, V.V. Strijov. Ordinal classification using Pareto fronts. Expert Systems with Applications, 42 (2015), pp. 5947-5953.
2. M.P. Kuznetsov and V.V. Strijov. Methods of expert estimations concordance for integral quality estimation // Expert Systems with Applications, 41(4):1988-1996, 2014.
3. A. M. Katrutsa, M. P. Kuznetsov, V. V. Strijov, K. V. Rudakov. Metric concentration search procedure using reduced matrix of pairwise distances // Intelligent Data Analysis, 19(5), 2015.
4. М. П. Кузнецов, В. В. Стрижов и М.М Медведникова. Алгоритм многоклассовой классификации объектов, описанных в ранговых шкалах. // Научно-технический вестник СПб ГПУ. Информатика. Телекоммуникации. Управление, 5, 2012.
5. М. М. Медведникова, В. В. Стрижов и М. П. Кузнецов. Алгоритм многоклассовой монотонной парето-классификации с выбором признаков. // Известия Тульского государственного университета, Естественные науки, 3:132-141, 2012.
6. В. В. Стрижов, М. П. Кузнецов и К. В. Рудаков. Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах. // Математическая биология и биоинформатика, 7(1):345-359, 2012.

Выступления с докладом

7. Математические методы распознавания образов ММРО-17, 2015. Комбинирование отношений порядка для восстановления предпочтения на наборе объектов.
8. 20th Conference of the International Federation of Operational Research Societies, Barcelona, Spain, 2014. Partial orders combining for object ranking problem.
9. European Conference on Operational Research. July 1-4, 2013. Rome, Italy. The IUCN Red List threatened speices categorization algorithm.
10. 25th European Conference on Operational Research. July 8-11, 2012. Vilnius, Lithuania. Rank-scaled Integral Indicators of Ecological Impact.
11. International Conference on Operational Research. August 30 to September 2, 2011. Zurich, Switzerland. Integral Indicators and Expert Estimations of Ecological Impact.