Дискретное квадратичное программирование с релаксацией при отборе признаков

Александр Катруца

Научный руководитель:

д.ф.-м.н. В. В. Стрижов

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

Москва
2016

Цели исследования

Цель работы

Предложить метод отбора признаков, учитывающий взаимное расположение признаков и целевого вектора.

Проблема

Методы отбора признаков дают избыточное подмножество мультикоррелирующих признаков.

Метод решения

Использование постановки задачи квадратичного программирования для получение оптимального подмножества признаков.

Постановка задачи

Пусть $\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n] \in \mathbb{R}^{m \times n}$ — матрица плана,
$\mathbf{y} \in \mathbb{R}^m$ — целевой вектор.

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^n} S(\mathbf{w}, \mathcal{A} | \mathbf{X}, \mathbf{y}, \mathbf{f}),$$

где S — функция ошибки, $\mathbf{f}$ — модель, $\mathcal{A}$ — множество активных признаков.
$S(\mathbf{w}, \mathcal{A} | \mathbf{X}, \mathbf{y}, \mathbf{f}) = \|\mathbf{f}(\mathbf{X}, \mathcal{A}, \mathbf{w}) - \mathbf{y}\|_2^2$ и $\mathbf{f}(\mathbf{X}, \mathcal{A}, \mathbf{w}) = \mathbf{X}_{\mathcal{A}}\mathbf{w}$,
$\mathbf{X}_{\mathcal{A}} \in \mathbb{R}^{m \times |\mathcal{A}|}$.

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} \mathcal{Q}(\mathcal{A} | \mathbf{X}, \mathbf{y}),$$

где $\mathcal{Q} : \mathcal{A} \to \mathbb{R}$ критерий качества подмножества индексов признаков $\mathcal{A} \subseteq \mathcal{J} = \{1, \ldots, n\}$.

$$\mathbf{a}^* = \arg\min_{\mathbf{a} \in \mathbb{B}^n} Q(\mathbf{a} | \mathbf{X}, \mathbf{y}),$$

где $Q : \mathbb{B}^n \to \mathbb{R}$.

Определение структуры выборки

Типы признаков:

- информативные — существенно влияют на точность приближения целевого вектора
- шумовые — не влияют на точность приближения целевого вектора
- мультиколлинеарные — существует линейная зависимость между признаками, снижают устойчивость модели

Задача квадратичного программирования

$$Q(\mathbf{a}|\mathbf{Q},\mathbf{b}) = \mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{a} - \mathbf{b}^\mathsf{T}\mathbf{a},$$

где матрица $\mathbf{Q} \in \mathbb{R}^{n\times n}$ и вектор $\mathbf{b} \in \mathbb{R}^n$.

- $\mathbf{Q} = [q_{ij}] = \mathrm{Sim}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)$ — похожесть между признаками i и j
- $\mathbf{b} = [b_i] = \mathrm{Rel}(\boldsymbol{\chi}_i)$ — релевантность признака i целевому вектору

Задача квадратичного программирования для отбора признаков:

$$\mathbf{a}^* = \arg\min_{\mathbf{a}\in\mathbb{B}^n} \mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{a} - \mathbf{b}^\mathsf{T}\mathbf{a}.$$

Примеры функций Sim и Rel

- Корреляция

$$q_{ij} = \left| \frac{\operatorname{cov}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)}{\sqrt{\operatorname{Var}(\boldsymbol{\chi}_i)\operatorname{Var}(\boldsymbol{\chi}_j)}} \right| \qquad b_i = \left| \frac{\operatorname{cov}(\boldsymbol{\chi}_i, \mathbf{y})}{\sqrt{\operatorname{Var}(\boldsymbol{\chi}_i)\operatorname{Var}(\mathbf{y})}} \right|$$

- Взаимная информация

$$I(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = \int \int p(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) \log \frac{p(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)}{p(\boldsymbol{\chi}_)p(\boldsymbol{\chi}_j)} d\boldsymbol{\chi}_i d\boldsymbol{\chi}_j.$$

Выпуклые релаксации

- Бинарные переменные $\rightarrow$ непрерывные переменные:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}\in[0,1]^n} \mathbf{z}^\mathsf{T}\mathbf{Q}\mathbf{z} - \mathbf{b}^\mathsf{T}\mathbf{z}$$

$$\text{s.t. } \|\mathbf{z}\|_1 \leq 1$$

- Невыпуклая целевая функция $\rightarrow$ выпуклая целевая функция
  - сдвиг спектра: $\hat{\mathbf{Q}} = \mathbf{Q} - \lambda_{\min}\mathbf{I}_n \succeq 0$, $\mathbf{I}_n$ единичная матрица, $\lambda_{\min}$ минимальное собственное значение матрицы $\mathbf{Q}$
  - полуопределённое программирование:

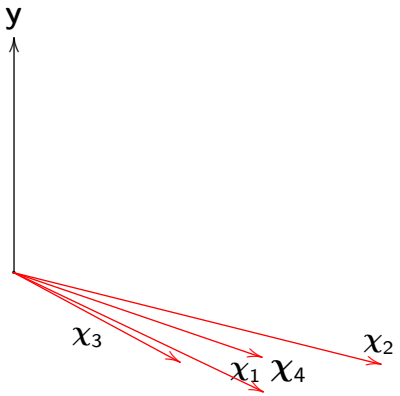
$$\mathbf{z}^* = \arg\min_{\mathbf{z}\in[0,1]^n, \mathbf{Z}\in\mathcal{S}_+} \operatorname{Tr}(\mathbf{Q}\mathbf{Z}) + \mathbf{b}^\mathsf{T}\mathbf{z}$$

$$\text{s.t. } \|\mathbf{z}\|_1 \leq 1,$$

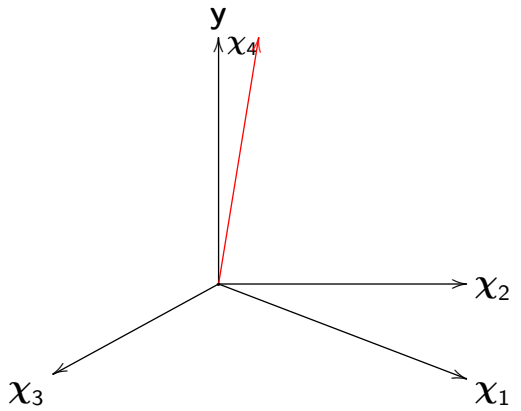
$$\begin{bmatrix} \mathbf{Z} & \mathbf{z} \\ \mathbf{z}^\mathsf{T} & 1 \end{bmatrix} \succeq 0.$$

$\boldsymbol{\chi}_j \in \mathcal{A}^* \Leftrightarrow z_j > \tau$, где τ — заданный порог.

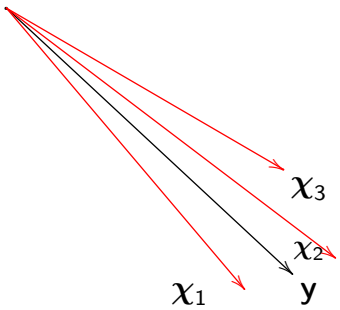
Тестовые выборки



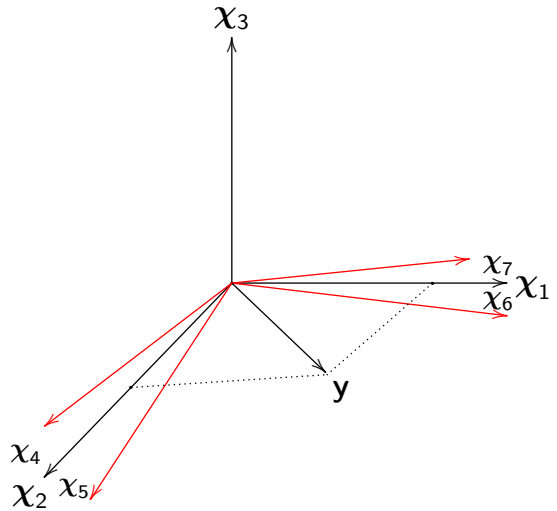
Неадекватная коррелирующая



Адекватная случайная



Адекватная избыточная



Адекватная коррелирующая

Вычислительный эксперимент

Цель эксперимента

Проверить работоспособность предложенного метода и сравнить с существующими методами отбора признаков

Название	*Формула*	*Значение*
VIF	$\mathrm{VIF} = \max\limits_{j \in \mathcal{A}} \frac{1}{1-R_j^2}$	Показатель мультиколлинеарности
Устойчивость	$R = \ln \frac{\lambda_{\min}}{\lambda_{\max}}$	Показатель устойчивости модели
Точность	$r = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$	Норма вектора остатков
Mallow's C_p	$C_p = \frac{r_p}{r} - m + 2p$	Баланс между точностью и сложностью
BIC	$\mathrm{BIC} = r + p \log m$	Баланс между точностью и сложностью

Сравнение с другими методами

Параметры: $m = 1000$, $n = 50$

Неадекватная коррелирующая выборка

Method	C_p	r	R	VIF	BIC
QP(ρ)	-997	—	—	—	—
LARS	-997	—	—	—	—
Genetic	-997	—	—	—	—
Lasso	-997	1	-6.57	16.6	310.48
Ridge	-997	1	-6.69	16.6	346.39
Stepwise	-997	1.68	-6.69	16.6	347.01
Elastic Net	-997	1	-6.58	16.6	310.48

Адекватная случайная выборка

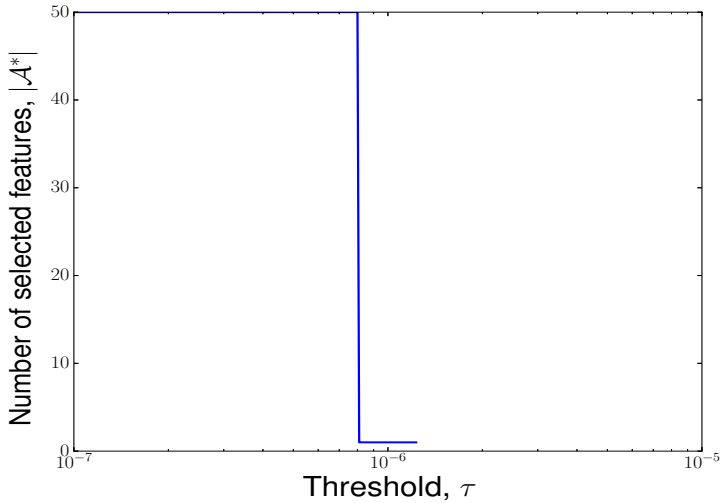
Method	C_p	r	R	VIF	BIC
QP(ρ)	-997	$1.2 \cdot 10^{-9}$	0	0.24	6.9
Lasso	$7 \cdot 10^{6}$	$8.50 \cdot 10^{-4}$	0	0.25	6.9
Elastic Net	$8.76 \cdot 10^{-4}$	$8.76 \cdot 10^{-4}$	0	0.25	6.9
Ridge	$7.97 \cdot 10^{9}$	0.97	0	0.25	7.88
LARS	-997	$1.3 \cdot 10^{-10}$	-0.78	0.32	8.29
Genetic	-997	$1.36 \cdot 10^{-10}$	-3.31	0.9	52.5
Stepwise	-997	$1.33 \cdot 10^{-10}$	-3.36	0.89	53.88

Сравнение с другими методами

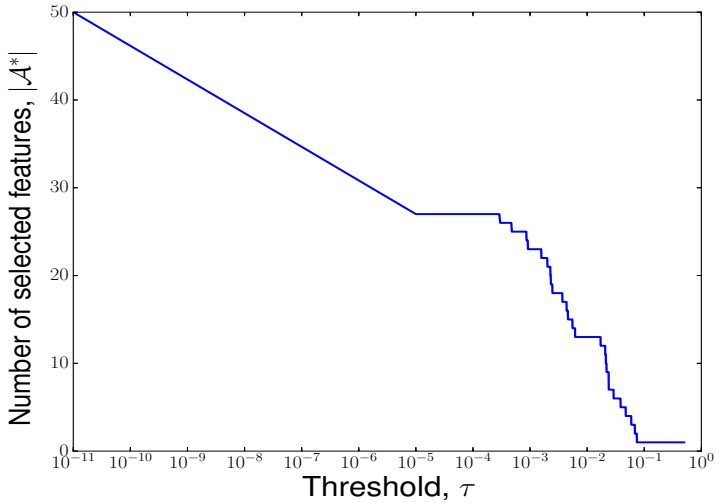
Адекватная избыточная выборка

Method	C_p	r	R	VIF	BIC
QP(ρ)	-997	$8.5 \cdot 10^{-11}$	0	0.25	6.9
Lasso	$5.16 \cdot 10^{8}$	$8.5 \cdot 10^{-4}$	0	0.24	6.9
Ridge	$5.9 \cdot 10^{11}$	0.97	-27.13	$2.9 \cdot 10^{9}$	346.36
Elastic Net	$5.16 \cdot 10^{8}$	$8.5 \cdot 10^{-4}$	-25.01	$2.5 \cdot 10^{9}$	41.45
Genetic	-997	$1.67 \cdot 10^{-12}$	-27.11	$2.87 \cdot 10^{9}$	345.39
Stepwise	-997	$1.73 \cdot 10^{-12}$	-27.13	$2.9 \cdot 10^{9}$	345.39
LARS	-997	$1.65 \cdot 10^{-12}$	-27.13	$2.9 \cdot 10^{9}$	345.39

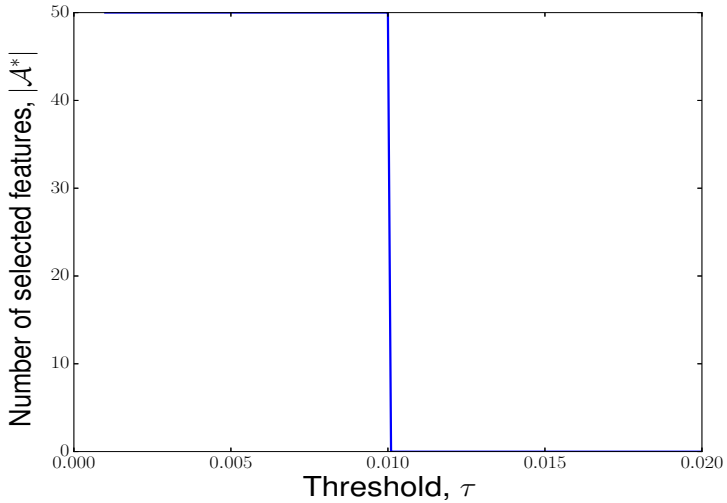
Результаты на тестовых выборках



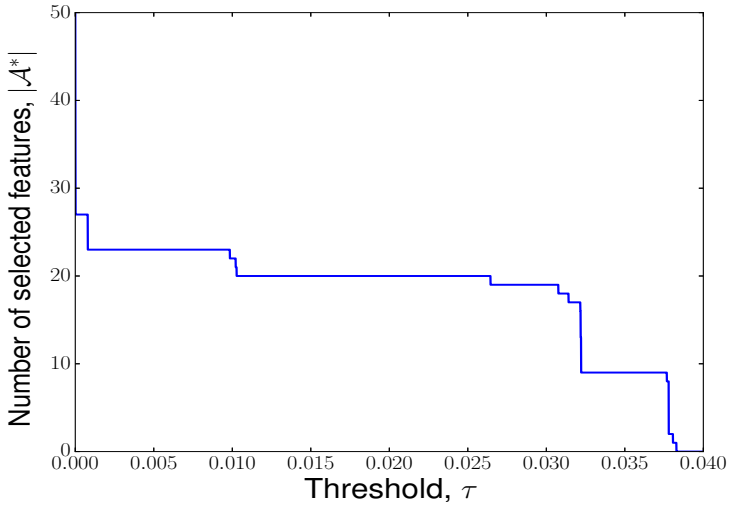
Неадекватная коррелирующая



Адекватная случайная

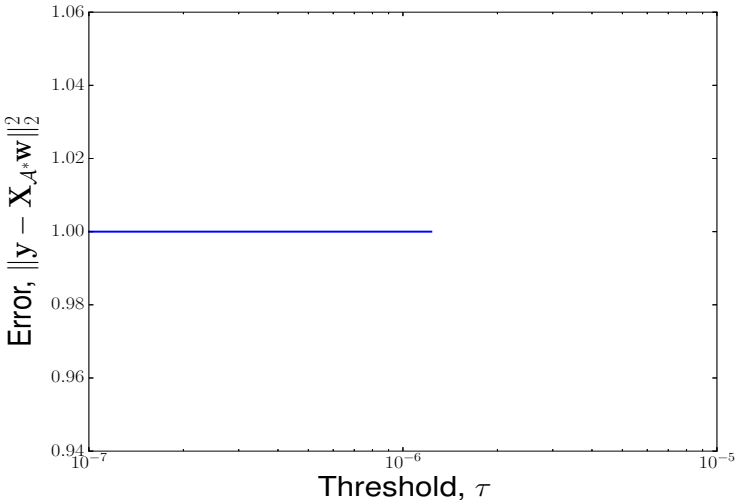


Адекватная избыточная

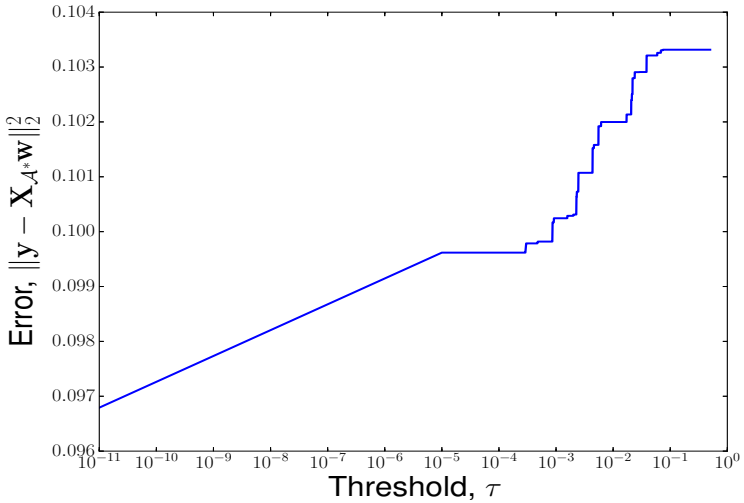


Адекватная коррелирующая

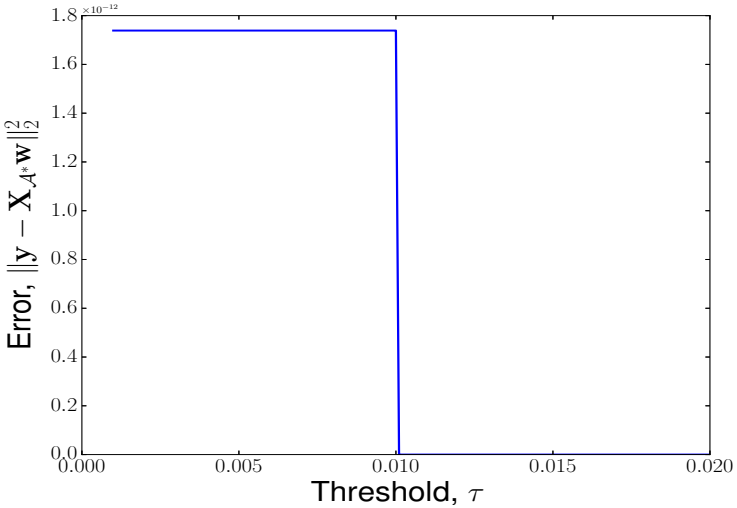
Результаты на тестовых выборках



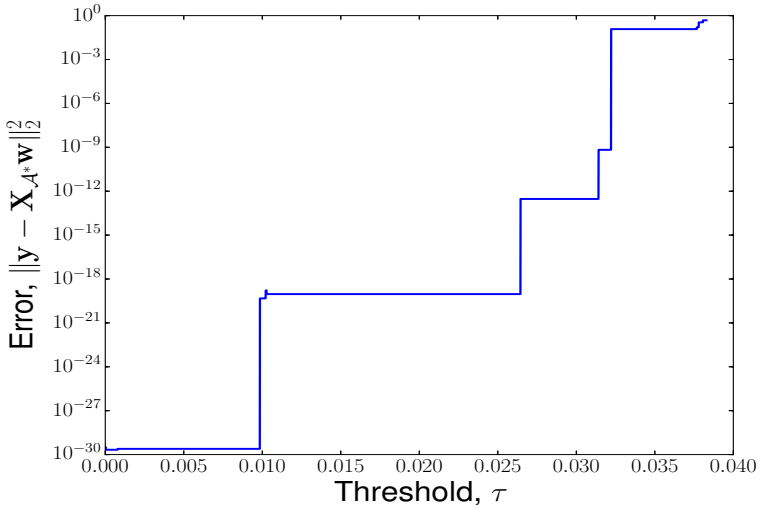
Неадекватная коррелирующая



Адекватная случайная

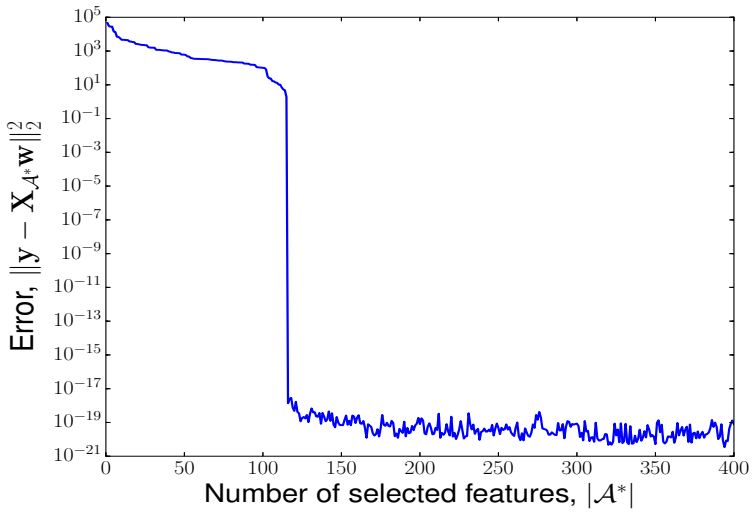


Адекватная избыточная

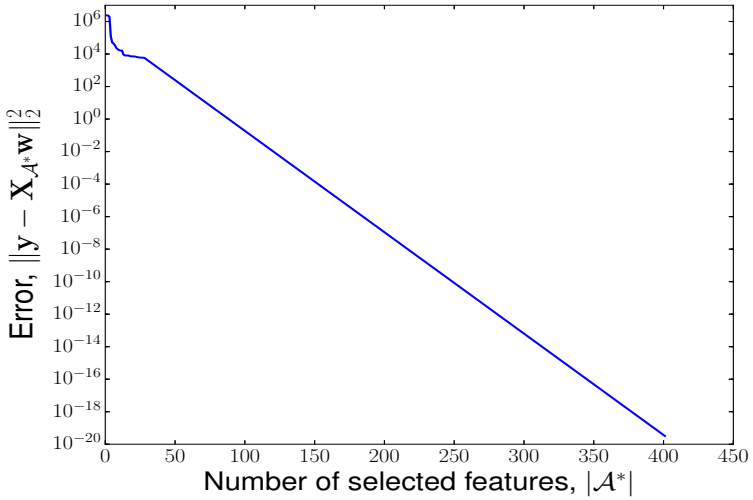


Адекватная коррелирующая

Результаты на реальных данных



Корреляция



Взаимная информация

Заключение

- Проблема выбора признаков сформулирована как задача квадратичного программирования
- Показана эффективность предложенного метода на тестовых выборках
- Проведено сравнение предложенного метода выбора признаков с другими методами

Публикации:

- *A.M. Katrutsa, V.V. Strijov* Stress test procedure for feature selection algorithms //Chemometrics and Intelligent Laboratory Systems. – 2015. – Т. 142. – С. 172-183. doi:10.1016/j.chemolab.2015.01.018
- *A.M. Katrutsa, M.P. Kuznetsov, V.V. Strijov* Metric concentration search procedure using reduced matrix of pairwise distances //Intelligent Data Analysis. – 2015. – Т. 19. – №. 5. – С. 1091-1108. doi:10.3233/IDA-150760
- *I.V. Oseledets, G.V. Ovchinnikov, A.M. Katrutsa* Fast, memory-efficient low-rank approximation of SimRank //Journal of Complex Networks. – 2016. doi:10.1093/comnet/cnw008
- *Нейчев Р. Г., Катруца А. М., Стрижов В. В.* Выбор оптимального набора признаков из мультикоррелирующего множества в задаче прогнозирования //Заводская лаборатория. Диагностика материалов. – 2016. – Т. 82. – №. 3. – С. 68-74.