Алгебраический подход к анализу данных и его приложения

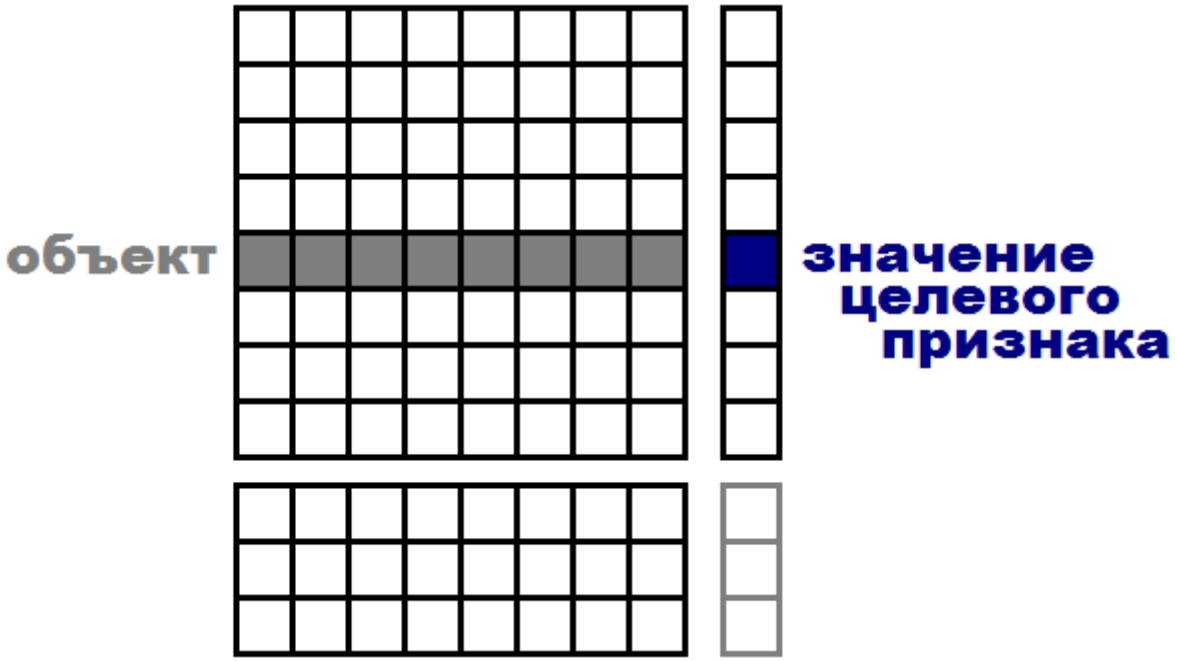
Александр Дьяконов

Московский государственный университет имени М.В. Ломоносова
Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

Область исследований

Анализ данных (Data Mining)

Машинное обучение (Machine Learning)



Область исследований

Задачи машинного обучения:

классификация (распознавание)
регрессия
прогнозирование

Примеры:

классификация спама
категоризация текстов
детектирование неисправностей
предсказание действий пользователей

Алгебраический подход к решению задач анализа данных (Ю.И. Журавлёв)

операции над алгоритмами
обоснование корректности полиномов над алгоритмами

Алгебраический подход к решению задач распознавания/классификации

$$I_0 \xrightarrow{B} \begin{matrix} \\ S_1 \\ \vdots \\ S_q \end{matrix} \begin{bmatrix} 1 & \dots & l \\ \gamma_{11} & \dots & \gamma_{1l} \\ \dots & \dots & \dots \\ \gamma_{q1} & \dots & \gamma_{ql} \end{bmatrix} \xrightarrow{C} \begin{matrix} \\ S_1 \\ \vdots \\ S_q \end{matrix} \begin{bmatrix} 1 & \dots & l \\ \alpha_{11} & \dots & \alpha_{1l} \\ \dots & \dots & \dots \\ \alpha_{q1} & \dots & \alpha_{ql} \end{bmatrix},$$

матрица оценок **матрица классификаций**

$$\alpha_{ij} \sim \text{«}S_i \in K_j\text{»}$$

$$A = B \cdot C,$$

B **– распознающий оператор,**

C **– решающее правило.**

модель АВО (алгоритмы вычисления оценок)

Алгебра над алгоритмами

Операции над распознающими операторами:

$$\Gamma[B_1 + B_2] = \Gamma[B_1] + \Gamma[B_2],$$

$$\Gamma[cB] = c\Gamma[B],$$

$$\Gamma[B_1 \cdot B_2] = \Gamma[B_1] \circ \Gamma[B_2].$$

Линейное замыкание $\mathbf{L}(B^*)$ множества B^*:

$$\mathbf{L}(B^*) = \{c_1 B_1 + \ldots + c_r B_r \mid r \in \{1,2,\ldots\},\ c_1,\ldots,c_r \in \mathbf{Q},\ B_1,\ldots,B_r \in B^*\}.$$

Алгебраическое замыкание k-й степени $\mathbf{U}^k(B^*)$:

$$\mathbf{U}^k(B^*) = \mathbf{L}(\{B_1 \cdot \ldots \cdot B_s \mid B_1,\ldots,B_s \in B^*,\ 1 \le s \le k\}).$$

Алгебраическое замыкание:

$$\mathbf{U}(B^*) = \bigcup_{k=1}^{\infty} \mathbf{U}^k(B^*).$$

Алгебра над алгоритмами

Определение. Модель распознающих операторов R^* называется корректной (относительно задачи распознавания), если

$$\forall \Gamma \in \mathbf{Q}^{q \times l} \ \exists B \in R^* : \Gamma[B] = \Gamma .$$

Теорема. Модель $U(B^*)$ корректна тогда и только тогда, когда выполнены первое и второе условия регулярности.

1–3 условия регулярности – Журавлёв Ю.И. (1977 г.)

1–2 условия регулярности – Докукин А.А. (2001 г.)

Основные результаты работы

1. Предложена теория систем эквивалентностей для описания и исследования алгебраических замыканий конечных степеней.

2. Получены новые критерии корректности алгебраического замыкания конечной степени и критерии разрешимости задач алгоритмами из этого замыкания.

3. Получена неулучшаемая в общем случае оценка степени корректного алгебраического замыкания модели АВО.

4. Исследовано пополнение линейного замыкания множества полиномов ограниченной степени над АВО операциями нормировки и деления.

5. Исследовано понятие корректности модели относительно семейства решающих правил.

6. Исследована k-сингулярность конечной системы точек (неполнота размерности пространства значений полиномов ограниченной степени над столбцами матрицы попарных l1-расстояний этой системы.

Новые критерии корректности алгебраического замыкания конечной степени и критерии разрешимости задач алгоритмами из этого замыкания

Построено множество операторов $\{B_{(i,j)}\}_{(i,j)\in QL}$, $QL=\{1,2,\ldots,q\}\times\{1,2,\ldots,l\}$**:**

Теорема. При $F_k\in\mathbf{F}^k$, $k\in\{1,2,\ldots\}$, **справедливо равенство**

$$\mathbf{L}(\{F_k(B_{(i,j)})\}_{(i,j)\in QL})=\mathbf{U}^k(B^*).$$

$\mathbf{F}^k$ **– полностью описано П.А. Карповичем (2010 г.)**

Полиномы Ю.И. Журавлёва (1977 г.)

$$\sum_i c_i B_i^{k(i)}$$

Получение любой матрицы оценки (для $\mathbf{U}^k(B^*)$**)**

$$\sum_{(a,b)\in QL} c_{(a,b)} B^k_{(a,b)}$$

$$\sum_{(a,b)\in QL} c_{(a,b)} F_k(B_{(a,b)})$$

Неулучшаемая в общем случае оценка степени корректного алгебраического замыкания модели АВО

$k \sim q \log q$ – **Ю.И. Журавлёв (1977 г.)**

$k \le q + l - 2$ – **В.Л. Матросов (1985 г.)**

$k \le m$ – **Т.В. Плохонина (1987 г.)**

$k \le \lfloor \log_2 ql \rfloor$ – **К.В. Рудаков (1989 г.)**

Точная оценка (общий случай):

$$k \le \lfloor \log_2 q \rfloor + \lfloor \log_2 l \rfloor$$

q – **число контрольных объектов,** l – **классов,** m – **эталонных объектов.**

Неулучшаемая в общем случае оценка степени корректного алгебраического замыкания модели АВО

Теорема. Модель $\mathrm{U}(B^*)$ корректна тогда и только тогда, когда корректна модель $\mathrm{U}^k(B^*)$, где $k \geq \lfloor \log_2 q \rfloor + \lfloor \log_2 l \rfloor$.

Теорема. Для любых натуральных параметров q и l, $q+l>2$, существует регулярная задача распознавания, в которой модель $\mathrm{U}^k(B^*)$ некорректна при $k < \lfloor \log_2 q \rfloor + \lfloor \log_2 l \rfloor$.

Также получены точные оценки для частных случаев.

Пополнение линейного замыкания множества полиномов ограниченной степени над АВО операциями нормировки и деления

Определение. **Стандартным замыканием относительно операций из** Op **множества** H^* **называется множество**

$$\mathbf{L}\,\mathrm{Op}\,\mathbf{L}(H^*) = \mathbf{L}\left(\bigcup_{F \in \mathrm{Op}} \{F(H) \mid H \in \mathbf{L}(H^*) \cap M^*(F)\} \cup H^* \right),$$

где $M^*(F)$ **– область определения операции** F.

Теорема. **Замыкание** $\mathrm{L}\,\mathrm{D}\,\mathrm{L}(B^*)$ **корректно тогда и только тогда, когда замыкание** $\mathrm{U}(B^*)$ **корректно, где** D **– непрерывная функция, отличная от полинома.**

$$\mathrm{D}(B_1 + 2B_2) - 3\mathrm{D}(-B_1 + B_2 - 5B_3) \qquad \mathrm{D}(B_1 + 2\mathrm{D}(B_2))$$

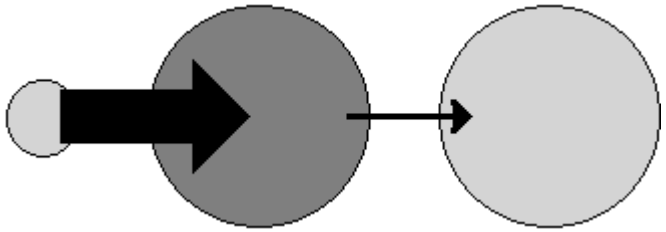
Корректности модели относительно семейства решающих правил

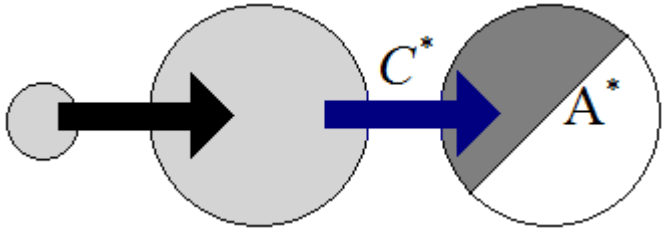
Пусть C^* – множество решающих правил,

A^* – множество классификаций ($q \times l$-матриц).

Определение. Модель R^* (распознающих операторов) называется C^*-A^*-корректной, если

$$\forall \hat{\mathrm{A}} \in \mathrm{A}^* \; \exists B \in R^*, \exists C \in C^* : C(\Gamma[B]) = \hat{\mathrm{A}}.$$



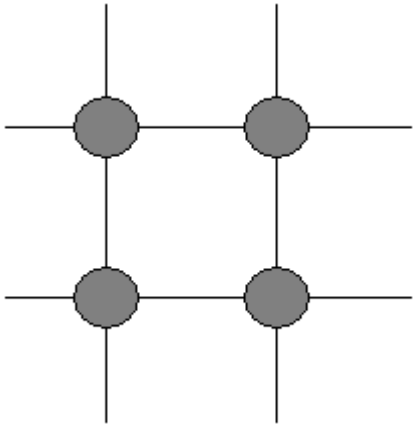


В большинстве случаев корректность эквивалентна C^*-A^*-корректности.

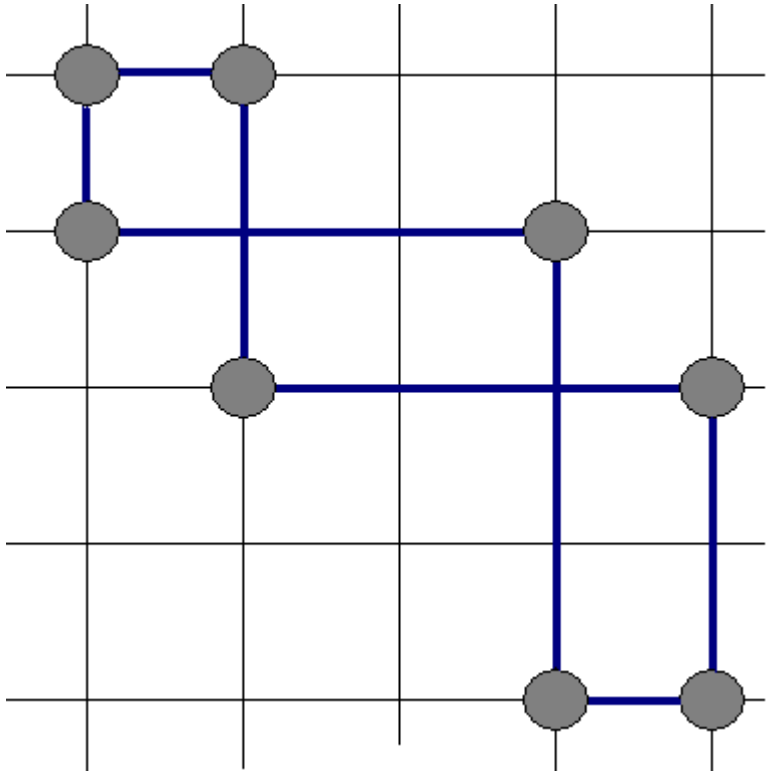
1-сингулярность конечной системы точек

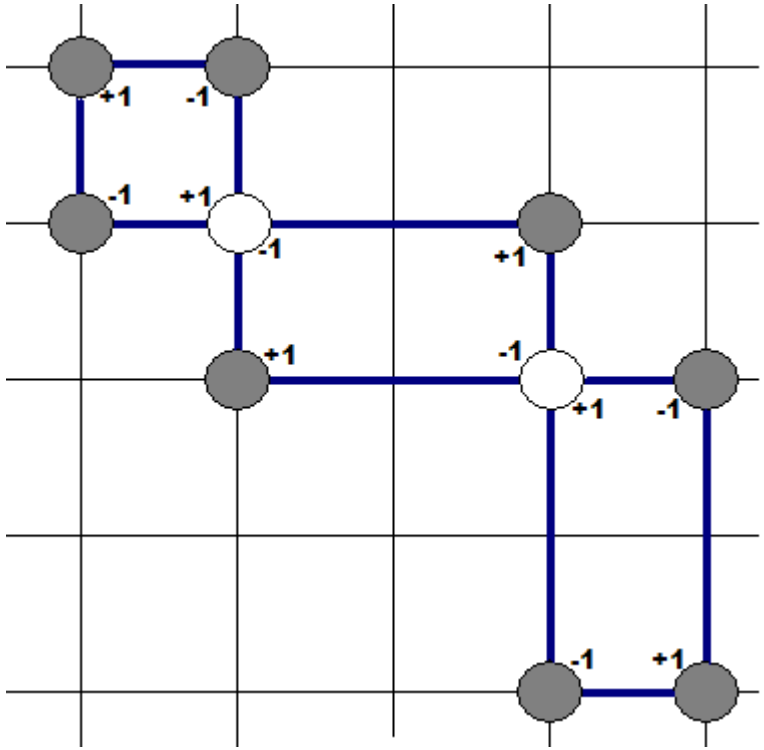
(вырожденность матрицы попарных l_1-расстояний этой системы)

Результаты И. Шёнберга: **Невырожденность матриц попарных расстояний для конечной системы** $\{\tilde{s}_i\}_{i=1}^{q}$ **различных точек пространства** $\mathbf{R}^m$ **и метрики** l_p, $1 < p \le 2$, $q > 1$.



$$\begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix}$$





1-сингулярность конечной системы точек

Теорема. **Система точек** $S=\{\tilde{s}_i\}_{i=1}^q$ **является** 1**-сингулярной тогда и только тогда, когда существует такое подмножество** $X\subseteq\{1,2,\ldots,q\}$**, что для любого преобразования** $g\in G$ **система точек** $\{g(\tilde{s}_i)\}_{i\in X}$ **не отделима от системы точек** $\{g(\tilde{s}_i)\}_{i\in\{1,2,\ldots,q\}\setminus X}$ **гиперплоскостью**

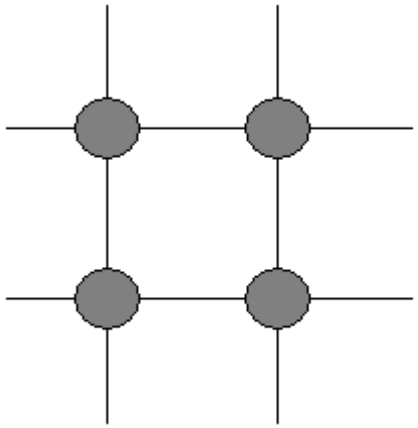
Теорема. **Система точек** $S=\{\tilde{s}_i\}_{i=1}^q$ **пространства** $\mathbf{R}^m$ **является** 1**-сингулярной тогда и только тогда, когда**

$$\exists(c_1,\ldots,c_q)^{\mathrm{T}}\in\mathbf{R}^q\setminus\{\tilde{0}\}\text{:}\quad\forall\tilde{s}\in\mathbf{R}^m\quad\sum_{i=1}^q c_i\rho(\tilde{s},\tilde{s}_i)=0,$$

где ρ **– метрика Хэмминга или** l_1**-метрика.**

k-сингулярность конечной системы точек

(неполнота размерности пространства значений полиномов ограниченной степени над столбцами матрицы попарных l_1-расстояний этой системы)



$$P_S = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix} \begin{matrix} 0 & 1 & 1 & 4 & 0 & 0 & 2 & 0 & 2 & 2 \\ 1 & 0 & 4 & 1 & 0 & 2 & 0 & 2 & 0 & 2 \\ 1 & 4 & 0 & 1 & 2 & 0 & 0 & 2 & 2 & 0 \\ 4 & 1 & 1 & 0 & 2 & 2 & 2 & 0 & 0 & 0 \end{matrix}$$

полиномы над столбцами

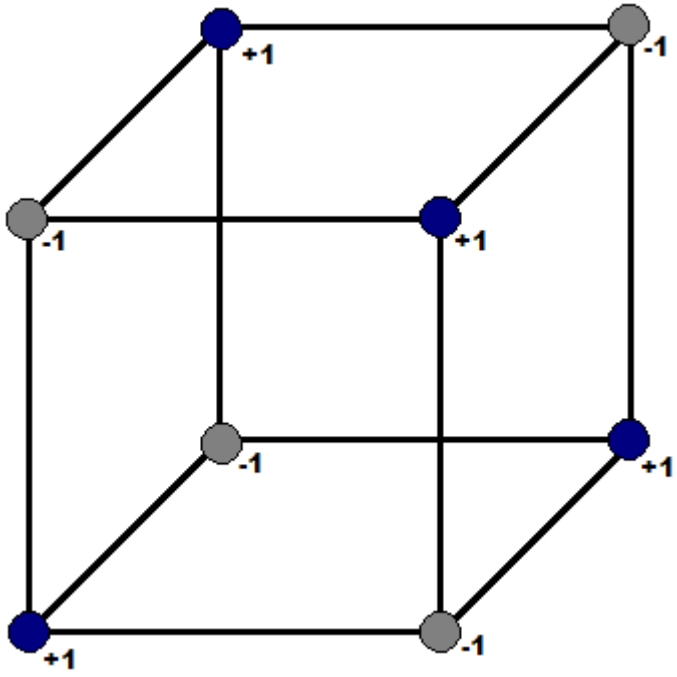
k-ранг

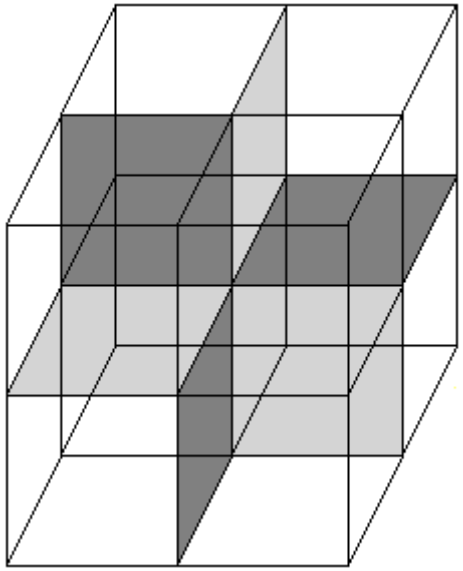
Определение. **Система точек** S **называется** k**-сингулярной, если размерность пространства** $\mathrm{U}^k[S]$ **меньше** q.

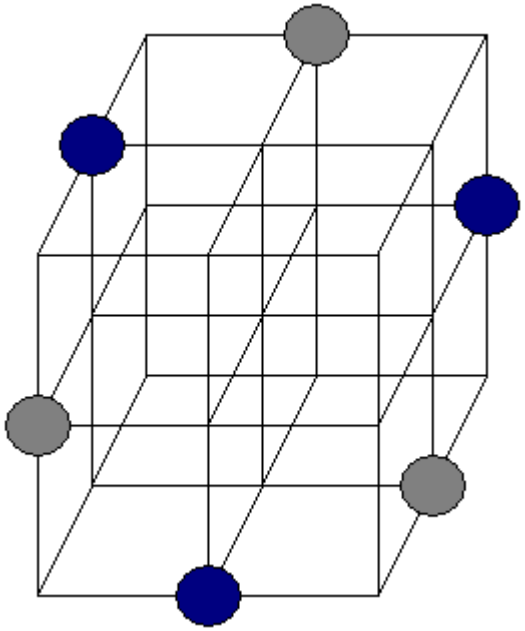
Критерии *k*-сингулярности

Теорема. **Система точек *k*-сингулярна тогда и только когда, на ней можно задать функцию, не представимую в виде суммы функций *k* переменных.**

Теорема. **Система точек *k*-сингулярна тогда и только тогда, когда содержит подсистему, которая является носителем суммы функций из множества D.**







Публикации

Дьяконов А.Г. Алгебра над алгоритмами вычисления оценок: минимальная степень корректного алгоритма // Ж. вычисл. матем. и матем. физ. – 2005. – Т.45. – №6. – С.1134–1145.

Дьяконов А.Г. Алгебра над алгоритмами вычисления оценок: монотонные решающие правила // Ж. вычисл. матем. и матем. физ. – 2005. – Т.45. – №10. – С.1893–1904.

Дьяконов А.Г. Алгебра над алгоритмами вычисления оценок: нормировка и деление // Ж. вычисл. матем. и матем. физ. – 2007. – Т.47. – №6. – С.1099–1109.

Дьяконов А.Г. Метрики алгебраических замыканий в задачах распознавания образов с двумя непересекающимися классами // Ж. вычисл. матем. и матем. физ. – 2008. – Т.48. – №5. – С.916–927.

Дьяконов А.Г. Критерии корректности алгебраических замыканий модели алгоритмов вычисления оценок // Докл. РАН. – 2008. – Т.420. – №6. – С.732–735.

Дьяконов А.Г. Алгебраические замыкания обобщенной модели алгоритмов вычисления оценок // Докл. РАН. – 2008. – Т.423. – №4. – С.461–464.

Дьяконов А.Г. Критерии вырожденности матрицы попарных l_1-расстояний и их обобщения // Докл. РАН. – 2009. – Т. 425. – №1. – С. 11–14.

Дьяконов А.Г. Алгебра над алгоритмами вычисления оценок: нормировка по отрезку // Ж. вычисл. матем. и матем. физ. – 2009. – Т.49. – №1. – С.200–208.

Дьяконов А.Г. Теория систем эквивалентностей для описания алгебраических замыканий обобщенной модели вычисления оценок // Журнал вычислительной математики и математической физики, 2010, Т. 50, №2. С.388-400.

Дьяконов А.Г. Теория систем эквивалентностей для описания алгебраических замыканий обобщенной модели вычисления оценок. II // Журнал вычислительной математики и математической физики, 2011, Т. 51, №3. С.529-544.

Дьяконов А.Г. Критерии вырожденности матрицы попарных l_1-расстояний и их обобщения // Известия РАН. Серия Математическая, 2012, Т.76:3, С. 93–110.

Прикладные задачи

год	соревнование	место	задача
2014	Greek Media Monitoring Multilabel Classification (WISE 2014)	1/121	Классификация текстов медиа-статей на большое число пересекающихся классов
2013	Олимпиада Викимарта	1/48	1. Прогнозирование вероятности заказа через колл-центр по действиям пользователя на сайте 2. Прогнозирование вероятности отказа пользователя от созданного заказа по описанию действий пользователя на сайте и описанию заказанных товаров 3. Прогнозирование вероятности ухода посетителя (прерывания сессии) с сайта по действиям пользователя на сайте
2011	VideoLectures.Net Recommender System Challenge (ECML/PKDD Discovery Challenge 2011)	1 / 62 1 / 22	Разработка алгоритма рекомендации лекций для просмотра на ресурсе VideoLectures.Net.
2011	dunnhumby's Shopper Challenge	1 / 287	Предсказывание визитов покупателей и сумм покупок для сети супермаркетов.
2008	Ford Classification Challenge	1 / 20	Классификация сигналов механизмов.

Алгоритм вычисления оценок (обобщённый)

– суперпозиция распознающего оператора (B) и решающего правила (C): $A = B \cdot C$.

Распознающий оператор B

Оценка принадлежности объекта S_i *к классу* K_j

$$\Gamma_{ij}[B] = \sum_{a,b=0,0}^{1,1} x_{ab}(j) \sum_{\Omega \in \Omega_A} \sum_{S^t \in \tilde{K}_j^a} w^t w(\Omega) \mathrm{B}_\Omega^{\tilde{e},b}(S^t, S_i),$$

где $w^t \in \mathbf{Q}^+$ при $t \in \{1,2,\ldots,m\}$ (*вес* t*-го объекта*), $w(\Omega) \in \mathbf{Q}^+$ при $\Omega \in \Omega_A$ (*вес учёта* Ω*-й близости*), $\Omega_A = \Omega_Z$, $\mathrm{B}_\Omega^{\tilde{e},b}(S^t, S_i)$ – *функция близости*:

$$\mathrm{B}_\Omega^{\tilde{e},1}(S^t, S_i) = 1 - \mathrm{B}_\Omega^{\tilde{e},0}(S^t, S_i) = \begin{cases} 1, & \rho_\Omega(S^t, S_i) \le \tilde{e}, \\ 0, & \rho_\Omega(S^t, S_i) \not\le \tilde{e}, \end{cases}$$ (или другой вид)

$\tilde{e}$ – параметры функции (из множества E_Z),

$$\tilde{K}_j^a = \begin{cases} \tilde{S}^m \cap K_j, & a = 1, \\ \tilde{S}^m \setminus K_j, & a = 0. \end{cases}$$