

Слабая вероятностная аксиоматика, оценки надёжности эмпирических предсказаний, расслоение и различность алгоритмов

Воронцов Константин Вячеславович
vokov@forecsys.ru, <http://www.ccas.ru/voron>

Вычислительный Центр РАН,
Москва, Вавилова 40, 119991

Интеллектуализация обработки информации,
ИОИ-2008, Крым, Алушта, 9–13 июня 2008

Содержание

- 1 Теория переобучения**
 - Задача оценивания вероятности переобучения
 - Теория Вапника-Червоненкиса и её развитие
 - Слабая (комбинаторная) вероятностная аксиоматика
- 2 Улучшение оценок и факторы завышенности**
 - Оценки Вапника-Червоненкиса в слабой аксиоматике
 - Факторы завышенности оценок ВЧ
 - Эксперимент
- 3 Расслоение и различность алгоритмов**
 - Переобучение в парах алгоритмов
 - Переобучение в цепочках алгоритмов
 - Выводы

Задача оценивания вероятности переобучения

Обучающая (наблюдаемая) выборка: $X^\ell = \{x_i\}_{i=1}^\ell \subset \mathbb{X}$.

Метод обучения $\mu: X^\ell \mapsto a$, где $a \in A$ — «алгоритм».

Индикатор ошибки алгоритма a на объекте $x \in \mathbb{X}$:

$$I(a, x), \quad I: A \times \mathbb{X} \rightarrow \{0, 1\}.$$

Вектор ошибок алгоритма a на выборке X^ℓ : $\vec{a} = (I(a, x_i))_{i=1}^\ell$

Частота ошибок алгоритма a на выборке X^ℓ

$$\nu(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} I(a, x_i).$$

Контрольная (скрытая) выборка: $X^k = \{x_i\}_{i=1}^k \subset \mathbb{X}$.

Переобученность:

— алгоритма a : $\delta(a, X^\ell, X^k) = \nu(a, X^k) - \nu(a, X^\ell)$.

— метода μ : $\delta(\mu, X^\ell, X^k) = \delta(\mu(X^\ell), X^\ell, X^k)$.

Задача: оценить сверху вероятность переобучения

$$P_{X^\ell, X^k} \{ \delta(\mu, X^\ell, X^k) > \varepsilon \} \leq \eta(\varepsilon), \quad \eta(\varepsilon) - ?$$

Оценки Вапника-Червоненкиса [1968]

Для любого множества алгоритмов A и любой меры P :

$$\begin{aligned} P_{X^L} \left\{ \delta(\mu, X^\ell, X^k) > \varepsilon \right\} &\leq \\ &\leq P_{X^L} \left\{ \sup_{a \in A} \delta(a, X^\ell, X^k) > \varepsilon \right\} \leq \\ &\leq \sum_a P_{X^L} \left\{ \delta(a, X^\ell, X^k) > \varepsilon \right\} \stackrel{\text{при } \ell = k}{\leq} \Delta^A(L) \cdot 1.5 e^{-\varepsilon^2 \ell}. \end{aligned}$$

$\Delta^A(L)$ — функция роста (shatter coefficient) множества A — максимальное (по всем X^L) число алгоритмов $a \in A$

с попарно различными векторами ошибок $\vec{a} = (I(a, x_i))_{i=1}^L$.

$\Delta^A(L) \leq 1.5 \frac{L^h}{h!}$, $h = \text{VCdim}(A)$ — ёмкость множества A .

Проблема: эта оценка крайне завышена и почти бесполезна для управления качеством алгоритма.

40 лет спустя: проблема остаётся открытой

- Равномерная сходимость [Вапник, Червоненкис, 1968]
- Theory of learnable (PAC-learning) [Valiant, 1982]
- Concentration inequalities [Talagrand, 1995]
- Connected function classes [Sill, 1995]
- Data-dependent bounds [Haussler, 1992; Bartlett, 1998;...]
- Self-bounding learning algorithms [Freund, 1998]
- PAC-Bayes bounds [McAllester, 1999; Langford, 2005]
- Microchoice bounds [Langford, Blum, 2001]
- Algorithmic luckiness [Herbrich, Williamson, 2002]
- Shell bounds [Langford, 2002]

John Langford. Quantitatively Tight Sample Complexity Bounds. PhD (Carnegie Mellon). 2002.

Причины завышенности: текущее понимание

- Это оценка «наихудшего случая»:
 - не учтены свойства выборки X^L и метода обучения μ ;
 - не учтена априорная информация о задаче.
- Это «union bound» $P(S_1 \cup \dots \cup S_\Delta) \leq P(S_1) + \dots + P(S_\Delta)$, где события есть $S_d = \{\delta(a_d, X^\ell, X^k) > \varepsilon\}$:
 - не учтено сходство алгоритмов.
- Не учтён эффект локализации: в каждой задаче «работает» только своя небольшая часть семейства.
- Экспоненциальный множитель $e^{-\varepsilon^2 \ell}$ завышен.
- **Цель:** сравнить причины завышенности количественно.
- **Конечная цель:** радикально улучшить оценку.
- **Проблема:** Вероятностная техника вывода оценок не позволяет контролировать их точность на каждом шаге.

Слабая вероятностная аксиоматика

- 1 $X^L = \{x_i\}_{i=1}^L$ — конечная неслучайная выборка объектов.
- 2 Все разбиения $X^L = X_n^\ell \cup X_n^k$, $n = 1, \dots, N$, $N = C_L^k$, имеют равные шансы реализоваться, $L = \ell + k$.
подвыборка X_n^ℓ — наблюдаемая;
подвыборка X_n^k — скрытая.

Переобученность при n -м разбиении: $\delta_n(\mu) \equiv \delta(\mu, X_n^\ell, X_n^k)$.

Вероятность определяется как доля разбиений:

$$P_n\{\delta_n(\mu) > \varepsilon\} = \frac{1}{N} \sum_{n=1}^N [\delta_n(\mu) > \varepsilon].$$

Замечание: понятие вероятности вводится без теории меры и без предельного перехода $L \rightarrow \infty$.

Преимущества слабой аксиоматики

- Не избыточна, не асимптотична.
- Допускает измерение вероятностей методом Монте-Карло (скользящий контроль):

$$\hat{P}_n\{\delta_n > \varepsilon\} = \frac{1}{|N'|} \sum_{n \in N'} [\delta_n > \varepsilon] \xrightarrow{N' \rightarrow N} P_n\{\delta_n > \varepsilon\}.$$

- Позволяет (если надо) вернуться к сильной аксиоматике: если $P_n\{\phi(X_n^\ell, X_n^k)\} \leq \eta(\varepsilon, X^L)$, то $P_{X^L}\{\phi(X^\ell, X^k)\} \leq E_{X^L}\eta(\varepsilon, X^L)$.
- Достаточна для доказательства фундаментальных фактов:
 - закон больших чисел: точная оценка скорости сходимости;
 - критерий Смирнова: точная оценка скорости сходимости;
 - многие непараметрические критерии;
 - оценки обобщающей способности (Вапника-Червоненкиса).

Открытый вопрос: какую часть математической статистики можно воспроизвести в рамках слабой аксиоматики?

Закон больших чисел в слабой аксиоматике

Пусть задан фиксированный алгоритм, $\nu(a, X^L) = m/L$.

Теорема (скорость сходимости частот в двух выборках)

Справедлива **точная** оценка

$$P_n\{\delta(a, X_n^\ell, X_n^k) \geq \varepsilon\} = H_L^{\ell, m}(s(\varepsilon)),$$

где $s(\varepsilon) = \lfloor \frac{\ell}{L}(m - \varepsilon k) \rfloor$;

$H_L^{\ell, m}(s) = \sum_{t=s_0}^s \frac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}$ — левый хвост гипергеометрического распределения,

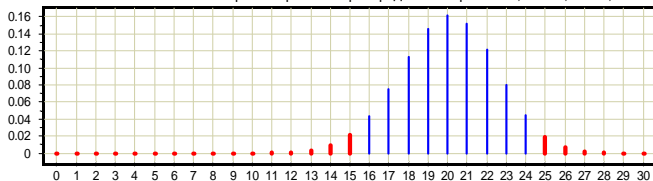
$$s_0 = \max\{0, m - k\}$$

Для получения численно точных оценок надо отказаться от использования верхних аппроксимаций г.г.р.

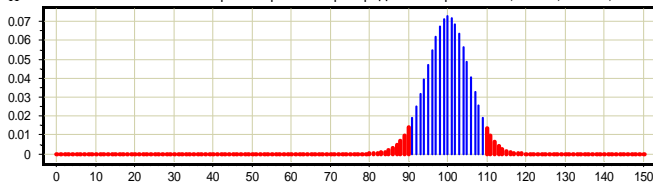
Закон больших чисел в слабой аксиоматике

Хвосты гипергеометрического распределения $H_L^{\ell,m}(s(\varepsilon))$:

H Гипергеометрическое распределение при L=300, k=100, m=30, eta=0.05



H Гипергеометрическое распределение при L=1500, k=500, m=150, eta=0.05



Оценки Вапника-Червоненкиса в слабой аксиоматике

Для любого метода μ и любой выборки X^ℓ :

$$\begin{aligned} Q_\varepsilon &= P_n \{ \delta(a_n, X_n^\ell, X_n^k) > \varepsilon \} \leq \\ &\leq \sum_{m=1}^L D_m \cdot H_L^{\ell, m}(s(\varepsilon)) \leq \\ &\leq \Delta_L^\ell \cdot \max_m H_L^{\ell, m}(s(\varepsilon)) \stackrel{\text{при } \ell = k}{\leq} \Delta^A(L) \cdot 1.5 e^{-\varepsilon^2 \ell}; \end{aligned}$$

$\Delta_L^\ell(\mu, X^L)$ — локальный коэффициент разнообразия (ЛКР, local shatter coefficient) множества алгоритмов $\{a_n = \mu(X_n^\ell) \mid n = 1, \dots, N\}$;

$D_m(\mu, X^L)$, $m = 0, \dots, L$ — профиль разнообразия — ЛКР множества алгоритмов, допускающих m ошибок на X^L : $\{a_n = \mu(X_n^\ell) \mid \nu(a_n, X^L) = \frac{m}{L}, n = 1, \dots, N\}$.

Оценки Вапника-Червоненкиса в слабой аксиоматике

В результате:

- Частично учтён эффект локализации.
- Избавились от экспоненциальной аппроксимации $1.5 e^{-\varepsilon^2 \ell}$.
- Учтено расслоение алгоритмов по числу ошибок m , но:
 - не ясно, как вычислять D_m ;
 - не ясно, даст ли это выигрыш.

Идея: измерить факторы завышенности в эксперименте

Определение

Локальный эффективный коэффициент разнообразия (ЛЭКР):

$$\hat{\Delta}_L^\ell(\varepsilon) = \frac{\hat{P}_n\{\delta(a_n, X_n^\ell, X_n^k) > \varepsilon\}}{H_L^{\ell, m}(s(\varepsilon))} = \frac{\hat{P}_n\{\delta(\mathbf{a}_n, X_n^\ell, X_n^k) > \varepsilon\}}{\hat{P}_n\{\delta(\mathbf{a}, X_n^\ell, X_n^k) > \varepsilon\}}.$$

Факторы завышенности оценок ВЧ

Степень завышенности оценки ВЧ раскладывается в виде:

$$\frac{\Delta^A(L) \cdot \frac{3}{2} e^{-\varepsilon^2 \ell}}{\hat{Q}_\varepsilon} = \underbrace{\frac{\Delta^A(L)}{\Delta_L^\ell}}_{r_1} \cdot \underbrace{\frac{\Delta_L^\ell}{\hat{\Delta}_L^\ell(\varepsilon)}}_{r_2(\varepsilon)} \cdot \underbrace{\frac{\hat{\Delta}_L^\ell(\varepsilon) \cdot H}{\hat{Q}_\varepsilon}}_{r_3(\varepsilon)} \cdot \underbrace{\frac{\frac{3}{2} e^{-\varepsilon^2 \ell}}{H}}_{r_4(\varepsilon)}$$

- $r_1 \geq 1$: пренебрежение эффектом локализации
- $r_2 \geq 1$: пренебрежение сходством алгоритмов («union bound» для выделения D_m как сомножителей)
- $r_3 \geq 1$: свёртка профиля разнообразия
- $r_4 \geq 1$: экспоненциальная аппроксимация гипергеометрического распределения $H = \max_m H_L^{\ell, m}(s(\varepsilon))$

Логические алгоритмы классификации

- *Закономерность* (правило) — предикат $\phi_y: X \rightarrow \{0, 1\}$, выделяющий преимущественно объекты класса y .
- *Взвешенное голосование* правил:

$$a(x) = \arg \max_{y \in Y} \sum_{t=1}^{T_y} w_y^t \phi_y^t(x),$$

где $\phi_y^t(x)$ — t -ое правило класса y , w_y^t — вес правила.

- *Метод обучения закономерностей* класса y :
 $\mu_y: X^\ell \mapsto \{\phi_y^t(x) \mid t = 1, \dots, T_y\}$.
- **Логические алгоритмы удобны для этого эксперимента:**
 - известна функция роста $\Delta^A(L)$;
 - легко оценить снизу ЛКР $\Delta_L^\ell(\mu, X^L)$;
 - легко оценить ЭЛКР $\hat{\Delta}_L^\ell(\varepsilon)$.

Экспериментальный стенд

- 7 задач классификации из репозитория UCI, $|\mathbb{Y}| = 2$
- 20×2 -кратный скользящий контроль, $\ell = k$
- Алгоритм Forecsys ScoringAce[®] [Кочедыков, Ивахненко,...]

Задача	L	n	C4.5	C5.0	RIPPER	SLIPPER	Forecsys
crx	690	15	15.5	14.0	15.2	15.7	14.3 ± 0.2
german	1000	20	27.0	28.3	28.7	27.2	28.5 ± 1.0
hepatitis	155	19	18.8	20.1	23.2	17.4	16.7 ± 1.7
horse-colic	300	25	16.0	15.3	16.3	15.0	16.4 ± 0.5
hypothyroid	3163	25	0.4	0.4	0.9	0.7	0.8 ± 0.04
liver	345	6	37.5	31.9	31.3	32.2	29.2 ± 1.6
promoters	106	57	18.1	22.7	19.0	18.9	12.0 ± 2.0

L — длина выборки; n — число признаков;

для алгоритмов указан процент ошибок на контроле.

Результаты

Степени завышенности при значении точности ε_0 ,
соответствующей надёжности $\hat{Q}_\varepsilon = 0.05$.

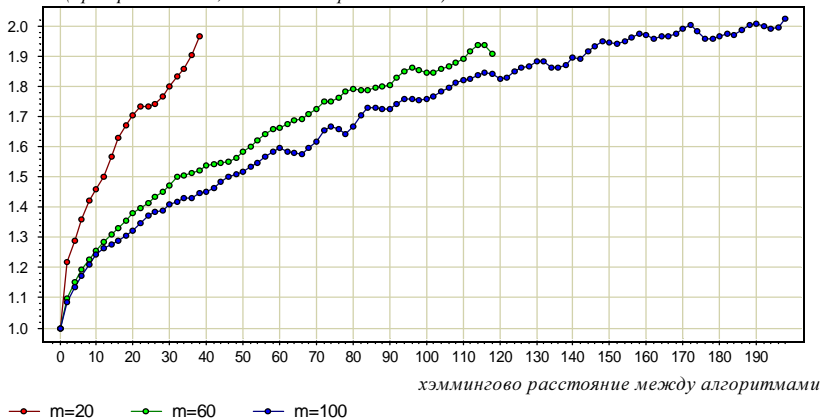
Задача	y	r_1	$r_2(\varepsilon_0)$	$r_3(\varepsilon_0)$	$r_4(\varepsilon_0)$	$\hat{\Delta}_L^\ell[\varepsilon_1, \varepsilon_2]$	$\hat{\Delta}_L^\ell(\varepsilon_0)$
crx	0	890	680	3.1	32.6	[10; 41]	24
	1	690	1700	1.6	11.6	[11; 180]	12
german	1	8 950	1500	1.7	10.9	[38; 530]	54
	2	37 000	9000	1.2	9.9	[1.0; 2.2]	1.9
hepatitis	0	23	280	13.4	9.5	[11; 148]	83
	1	55	680	2.4	22.5	[12; 27]	15
horse-colic	1	72	4500	2.1	7.2	[2; 9]	7
	2	140	3400	3.6	7.3	[3; 6]	6
hypothyroid	0	61 000	400	32.2	16.5	[3; 220]	21
	1	153 000	460	3.8	28.7	[2; 44]	30
promoters	0	94	340	5.9	9.8	[36; 230]	72
	1	150	790	3.4	6.9	[9; 22]	18

Выводы

- Локальный эффективный коэффициент разнообразия $\hat{\Delta}_L^\ell$ имеет порядок не выше 10^2 .
- «Эффективная локальная ёмкость» ≤ 1 .
- Оценки сложности Δ такого порядка пока не известны.
- Бессмысленно заниматься оцениванием профиля D_m .
- Основные пути улучшения оценок:
Локализация + Учёт сходства алгоритмов.

Vorontsov K. V. Combinatorial Probability and Generalization Bounds
Tightness // Pattern Recognition and Image Analysis. — 2008? (в печати).//
Русский «исходник»: www.ccas.ru/voron

Пара алгоритмов

Переобучение при выборе из пары алгоритмов ($L=200$, $k=100$)*ЭЛКР (при $\epsilon=0.05$, по $N=10000$ разбиениям)*

При увеличении различности переобучение увеличивается.

Имитационный эксперимент

$D = 1000$ алгоритмов, заданных векторами ошибок;

$\ell = k = 100$ — длина обучения и контроля;

$m = 10$, $m = 50$ — уровень ошибок 5%, 25%;

$\varepsilon = 0.05$ — порог переобученности;

$N' = 1000$ разбиений по методу Монте-Карло.

Генерируется бинарная $L \times D$ -матрица векторов ошибок:

Пример:

1	1	0	0	0	1	1	1	1	1	1	...
0	0	0	0	1	1	1	1	1	1	0	...
0	0	0	0	0	0	0	0	0	0	0	...
0	0	0	1	1	1	1	1	0	0	0	...
0	0	0	0	0	0	0	1	1	1	1	...
0	0	0	0	0	0	0	0	0	0	0	...
0	0	0	0	0	0	0	0	0	0	0	...
0	1	1	1	1	1	0	0	0	1	1	...
.

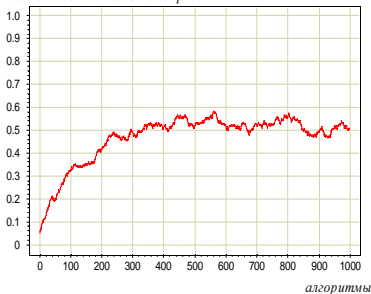
Цепочка алгоритмов

Цепочка — это последовательность векторов ошибок, в которой каждый следующий отличается от предыдущего на 1 объекте.

Два крайних случая цепочек:

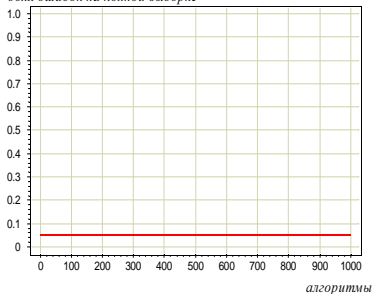
Цепочка 1000 алгоритмов с расслоением по m
($L=200, k=100, m_{\min}=10$)

доля ошибок на полной выборке



Цепочка 1000 алгоритмов с одинаковым m
($L=200, k=100, m=10$)

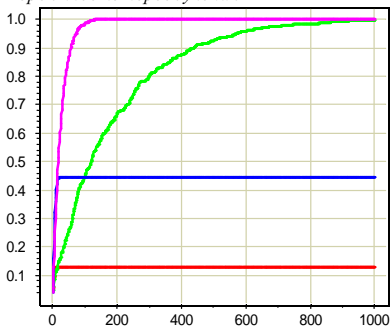
доля ошибок на полной выборке



Цепочки и не-цепочки, с расслоением и без (при 5% ошибок)

Наборы 1000 алгоритмов ($L=200$, $k=100$, $m_{\min}=10$, $\epsilon=0.05$, $N=1000$)

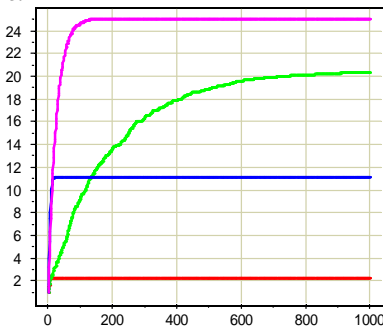
Вероятность переобучения



число алгоритмов в цепочке

—●— цепочка с расслоением по m —●— случайные с расслоением по m —●— цепочка с одинаковым m —●— случайные с одинаковым m

ЭЛКР



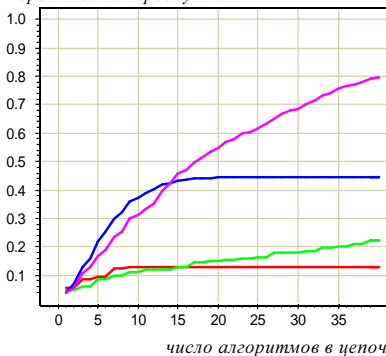
число алгоритмов в цепочке

Если расслоение есть, и уровень ошибок мал ($m = 5\%$),
то вероятность переобучения не достигает 1 с ростом D .

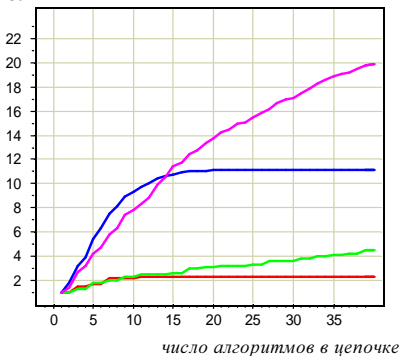
Цепочки и не-цепочки, с расслоением и без (5%, увеличено)

Наборы 1000 алгоритмов ($L=200$, $k=100$, $m_{\min}=10$, $\epsilon=0.05$, $N=1000$)

Вероятность переобучения



ЭЛКР

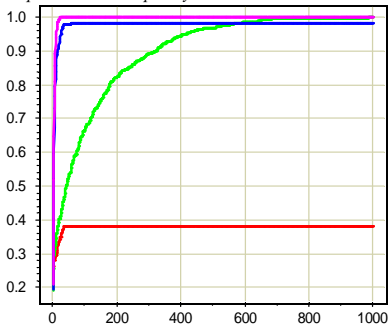
—●— цепочка с расслоением по m —●— случайные с расслоением по m —●— цепочка с одинаковым m —●— случайные с одинаковым m

По Вапнику $\hat{\Delta}(D) = D$. Это наблюдается только при малых D , и только для случайных векторов ошибок (не-цепочек).

Цепочки и не-цепочки, с расслоением и без (при 25% ошибок)

Наборы 1000 алгоритмов ($L=200$, $k=100$, $m_{\min}=50$, $\epsilon=0.05$, $N=1000$)

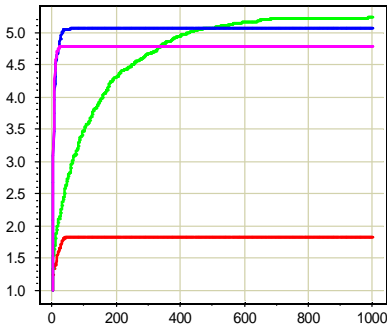
Вероятность переобучения



число алгоритмов в цепочке

—●— цепочка с расслоением по m —●— случайные с расслоением по m —●— цепочка с одинаковым m —●— случайные с одинаковым m

ЭЛКР



число алгоритмов в цепочке

При увеличении уровня ошибок до $m = 25\%$ только цепочка с расслоением не переобучена.

Выводы

- Переобучение возникает в результате выбора алгоритма (хотя бы из двух!), лучшего на конечной выборке.
- Функционал равномерной сходимости (по Вапнику) не позволяет учесть эффекты расслоения и различности.
- Зависимость $\hat{\Delta}(D)$ всегда «выходит на насыщение» (тогда как по Вапнику $\Delta = D$).
- В цепочках эта зависимость растёт на порядок медленнее.
- В расслоениях эта зависимость «насыщается», не достигая 1.
- При больших D только наличие цепочек и расслоения вместе спасает от сильного переобучения.
- **Есть повод для оптимизма: в практических задачах как раз есть и цепочки, и расслоения.**

Спасибо за внимание!

Предлагается открыть *виртуальный семинар*
по вопросам обобщающей способности
на страницах вики-ресурса
www.MachineLearning.ru

Участник: Vokov (К. В. Воронцов)
vokov@forecsys.ru

пятница, 13 июня, в конце утреннего заседания \approx 12:00

Заглавная страница - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.machinelearning.ru/wiki/index.php?title=%D0%97%D0%B0%D0%B3%D0%BE%D0%B0%D0%B2%D0%B0%D0%B0%D1%8F_%D1%81%D1%82%D1%80%D0%BE%... Go Links

Vokov моя страница обсуждения настройки список наблюдения мой вклад завершение сеанса

статья обсуждение править история удалить переименовать снять защиту не следить

MachineLearning.Ru

Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.

Сейчас ресурс содержит 105 статей на русском языке.

Классификация	Обработка и анализ текстов
Прогнозирование	Анализ и понимание изображений
Регрессионный анализ	Извлечение знаний из баз данных
Прикладная статистика	Прикладные задачи анализа данных
Распознавание образов	Прикладные системы анализа данных
Обработка сигналов	Все направления

Концепция Инструктаж Все статьи Непубликуемые статьи Полезные ссылки Частые вопросы Справка

Цели Ресурса

- Сконцентрировать информацию о достижениях ведущих российских научных школ в области машинного обучения и интеллектуального анализа данных.
- Способствовать обмену опытом, накоплению и распространению научных знаний в этой области.
- Предоставить площадку для виртуальных научных семинаров и обсуждений.
- Предоставить доступ к распределенной системе тестирования алгоритмов классификации и прогнозирования.

Основные принципы

Ресурс строится по принципам Википедии — свободной энциклопедии.

Содержимое Ресурса создается всеми его пользователями

Последние новости

- С 15 по 20 сентября 2008 года в Нижнем Новгороде будет проведена международная конференция РОАИ-9-2008. Подробная информация — на сайте конференции [df](#).
- С 9 по 14 июня 2008 года в г. Алуште согласно планам научных мероприятий академий наук России, Украины и Беларуси будет проведена Международная конференция «Интеллектуализация обработки информации». В рамках конференции планируется презентация и обсуждение Ресурса.
- 20 мая 2008 года — Количество страниц в базе данных ресурса достигло 500 (из них 97 статей). Создано пространство имён «Публикации», в тестовом режиме заведено 4 публикации.

Все новости

Done Internet