О некоторых задачах и методах интеллектуального анализа данных

д.ф.-м.н. Константин Вячеславович Воронцов
(voron@forecsys.ru),
к.ф.-м.н. В. В. Стрижов

кафедра «Интеллектуальные системы» ФУПМ МФТИ

специализация «Интеллектуальный анализ данных»
(отдел «Интеллектуальные системы» ВЦ РАН + ЗАО «Форексис»)

специализация «Информационный поиск и машинное обучение»
(отдел «Интеллектуальные системы» ВЦ РАН + ШАД Яндекс)

Содержание

Задача обучения по прецедентам (машинного обучения)

$\mathbb{X}$ — объекты; $\mathbb{Y}$ — ответы (классы, прогнозы);
$y^*: \mathbb{X} \to \mathbb{Y}$ — неизвестная зависимость.

Дано: $x_i = (x_i^1, \ldots, x_i^n)$ — *обучающие объекты*
с известными ответами $y_i = y^*(x)$, $i = 1, \ldots, \ell$:

$$\begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots \\ x_\ell^1 & \ldots & x_\ell^n \end{pmatrix} \xrightarrow{y^*} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: алгоритм $a: \mathbb{X} \to \mathbb{Y}$, способный давать правильные ответы на новых объектах $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \ldots & \tilde{x}_1^n \\ \ldots & \ldots & \ldots \\ \tilde{x}_k^1 & \ldots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_k) \end{pmatrix}$$

Примеры прикладных задач обучения по прецедентам

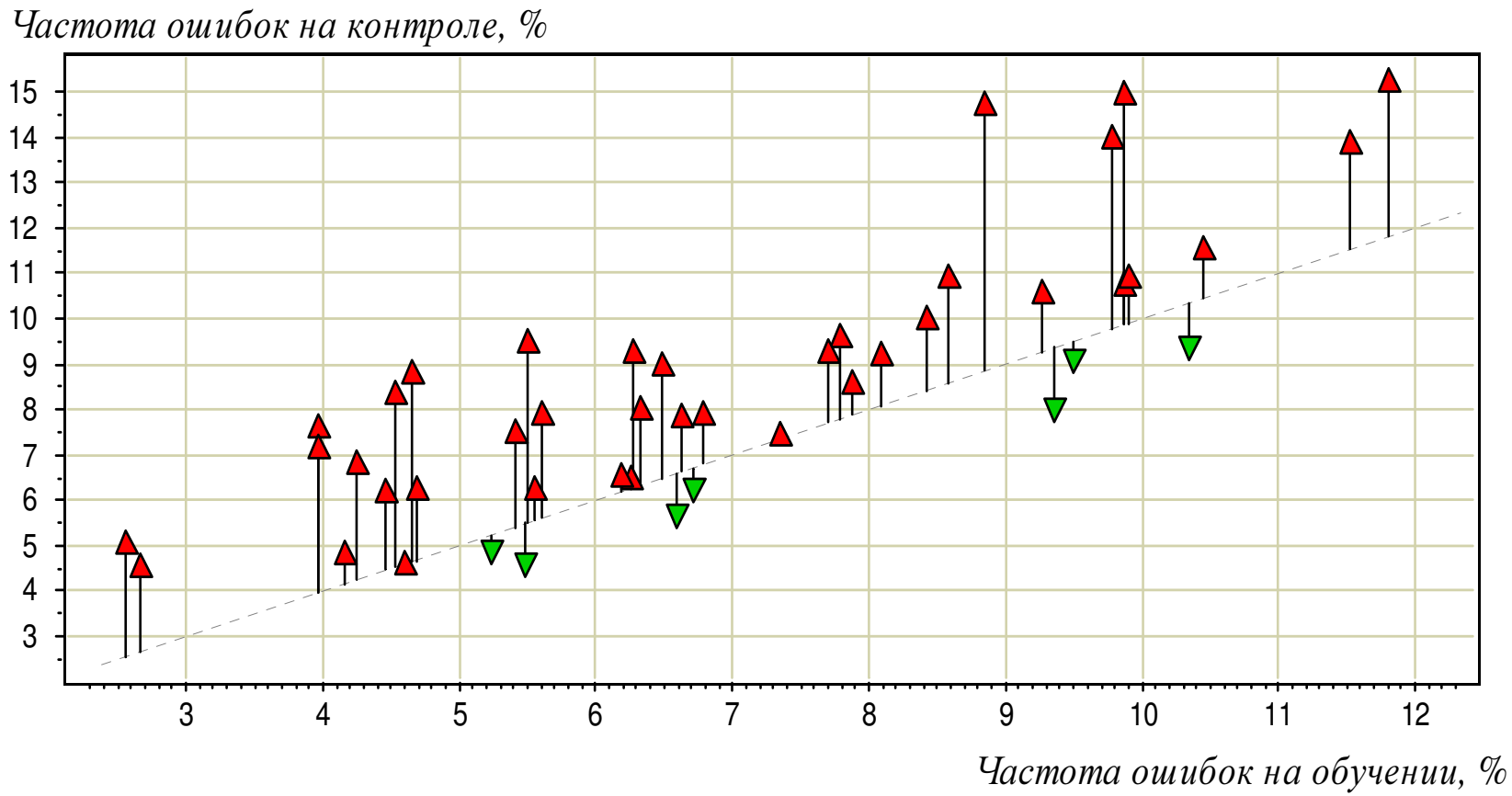
- Распознавание, классификация, принятие решений ($|\mathbb{Y}| < \infty$):
  - x — пациент; y — диагноз, рекомендуемая терапия;
  - x — заёмщик; y — вероятность дефолта;
  - x — абонент; y — вероятность ухода к другому оператору;
  - x — текстовое сообщение; y — спам / не спам;
  - x — документ; y — категория в рубрикаторе;
  - x — фрагмент белка; y — тип вторичной структуры;
  - x — фрагмент ДНК; y — функция: промотор / ген;
  - x — фотопортрет; y — идентификатор личности;
- Регрессия и прогнозирование ($\mathbb{Y} = \mathbb{R}$ или $\mathbb{R}^m$):
  - x — история продаж; y — прогноз объёма продаж;
  - x — пара $\langle$клиент,товар$\rangle$; y — рейтинг товара;
  - x — параметры технолог. процесса; y — свойство продукции;
  - x — структура хим. соединения; y — его свойство;
  - x — характеристики недвижимости; y — цена;

Особенности реальных задач обучения по прецедентам

- Особенности исходных данных:
  - неполнота данных (пропуски);
  - неточность данных (погрешности, выбросы);
  - разнородность (сложные «сырые» данные);
  - несбалансированность классов;
  - малые выборки;
  - сверхбольшие выборки;
  - потоковые данные;
  - нестандартные критерии качества;
  - наличие дополнительной непрецедентной информации.
- Требования к методам восстановления зависимостей:
  - обобщающая способность (минимум переобучения);
  - вычислительная эффективность;
  - простота и интерпретируемость модели;
  - визуализация и контроль промежуточных данных;
  - динамическое дообучение по потоковым данным;

Пример переобучения. Реальная задача классификации

Задача предсказания отдалённого результата хирургического лечения атеросклероза, $L = 98$. Точки — различные алгоритмы.



Модель принятия решений по неполной информации

$\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов;
$A = \{a_1, \ldots, a_D\}$ — конечное множество *алгоритмов*;
$I(a, x) = [$алгоритм a ошибается на объекте $x]$;

$L \times D$-*матрица ошибок* с попарно различными столбцами:

	a_1	a_2	a_3	a_4	a_5	a_6	$\cdots$	a_D	
x_1	1	1	0	**0**	0	1	$\cdots$	1	X — наблюдаемая
$\ldots$	0	0	0	**0**	1	1	$\cdots$	1	(обучающая) выборка
x_ℓ	0	0	1	**0**	0	0	$\cdots$	0	длины ℓ
$x_{\ell+1}$	0	0	0	**1**	1	1	$\cdots$	0	$\bar{X}$ — скрытая
$\ldots$	0	0	0	**1**	0	0	$\cdots$	1	(контрольная) выборка
x_L	0	1	1	**1**	1	1	$\cdots$	0	длины $k = L - \ell$

$n(a, X) = \sum_{x \in X} I(a, x)$ — *число ошибок* $a \in A$ на выборке $X \subset \mathbb{X}$;
$\nu(a, X) = \frac{1}{|X|} n(a, X)$ — *частота ошибок* a на выборке X;

Формализация понятия «обобщающая способность»

Обучение методом *минимизации эмпирического риска*:

$$\mu(X) = \arg\min_{a \in A} \nu(a, X).$$

Основная вероятностная аксиома

Все разбиения $X \sqcup \bar{X} = \mathbb{X}$ равновероятны, $|X| = \ell$, $|\bar{X}| = k$.

В этом случае $\mathsf{P} \equiv \mathsf{E} \equiv \frac{1}{C_L^\ell} \sum_{X \subset \mathbb{X}}$ — доля разбиений выборки.

Функционалы обобщающей способности

- ожидаемая частота ошибок на контроле:

$$\mathrm{CCV}(\mu, \mathbb{X}) = \mathsf{E}\, \nu\big(\mu(X), \bar{X}\big).$$

- вероятность переобучения:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu\big(\mu(X), \bar{X}\big) - \nu\big(\mu(X), X\big) \geqslant \varepsilon\big].$$

Теория Вапника–Червоненкиса

Теорема (Вапник, Червоненкис, 1971)

Для любых $\mathbb{X}$, A, μ и $\varepsilon \in [0,1]$, при $\ell = k$

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant |A| \cdot \tfrac{3}{2} \exp\bigl(-\varepsilon^2 \ell\bigr).$$

Основной вывод: переобучение может увеличиваться с ростом сложности семейства $|A|$.

Проблема завышенности:

- эта оценка завышена в 10^8–10^{11} раз;
- что приводит к оценкам длины обучения $\ell = 10^6$–10^{10}.

Причина завышенности — это оценка «худшего случая»:

- она зависит только от размеров матрицы ошибок $L \times D$;
- не зависит от её содержимого $I(a,x)$, выборки $\mathbb{X}$, метода μ.

Комбинаторная теория переобучения

Определим бинарные отношения на множестве алгоритмов A:

частичный порядок $a \leqslant b$: $I(a,x) \leqslant I(b,x)$ для всех $x \in \mathbb{X}$;

предшествование $a \prec b$: $a \leqslant b$ и $\|b-a\| = 1$.

Определение

Граф расслоения–связности $\langle A, E \rangle$:

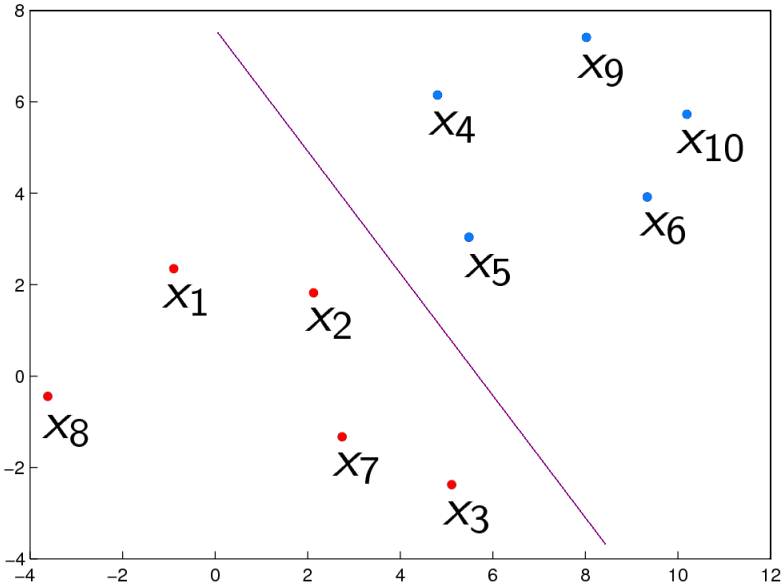
A — множество попарно различных векторов ошибок;

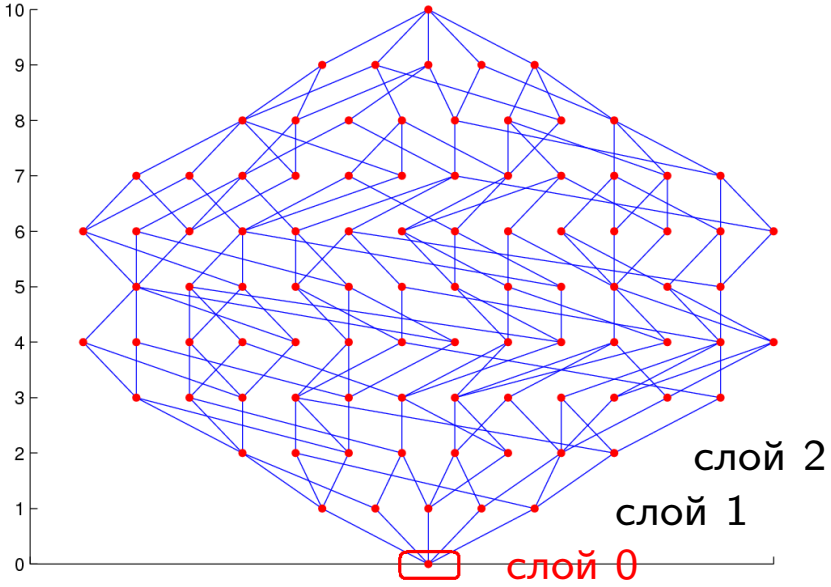
$E = \{(a,b)\colon a \prec b\}$.

Свойства графа расслоения–связности:

- это подграф графа Хáссе отношения порядка $\leqslant$ на A;
- каждому ребру (a,b) соответствует объект $x_{ab} \in \mathbb{X}$, такой, что $I(a,x_{ab}) = 0, \quad I(b,x_{ab}) = 1$;
- граф является многодольным со слоями $A_m = \{a \in A\colon n(a,\mathbb{X}) = m\}, \quad m = 0,\ldots,L$;

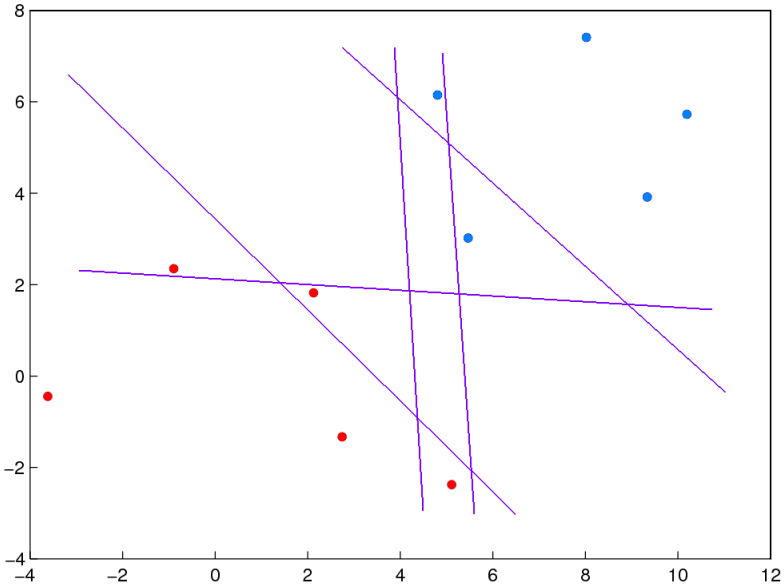
Пример. Семейство линейных алгоритмов классификации

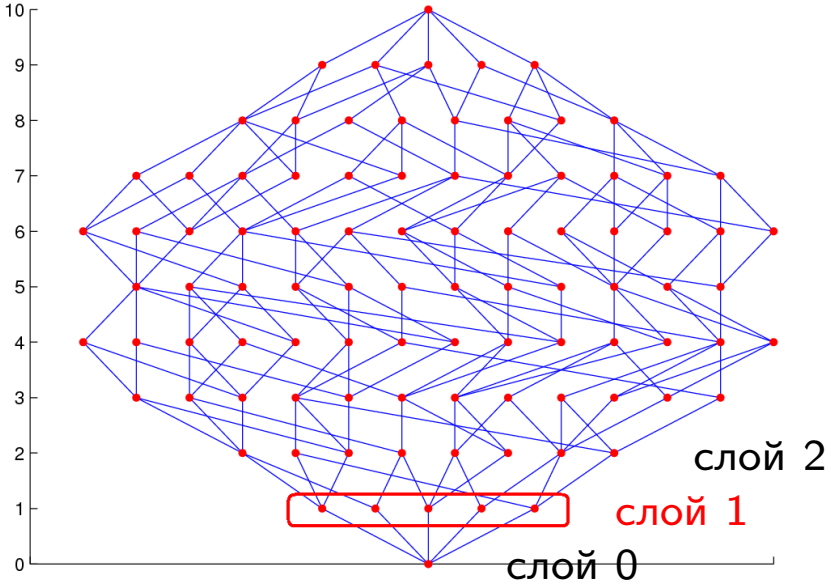




	слой 0
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

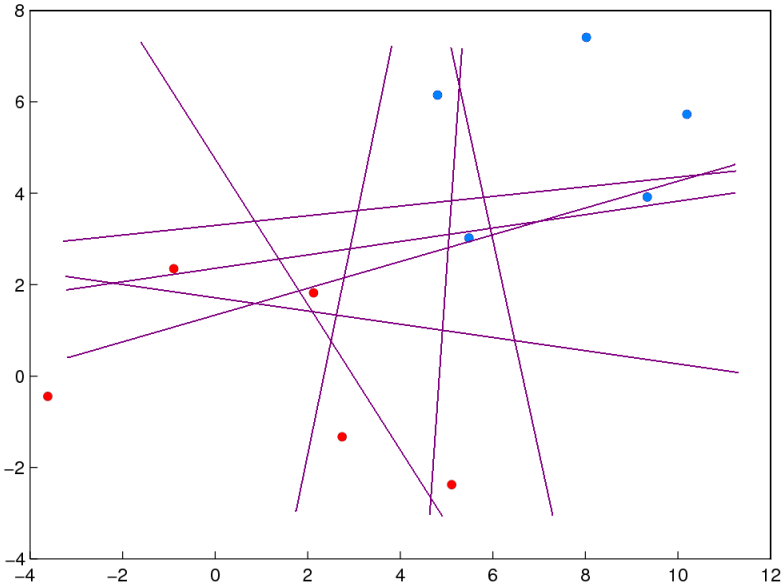
Пример. Семейство линейных алгоритмов классификации

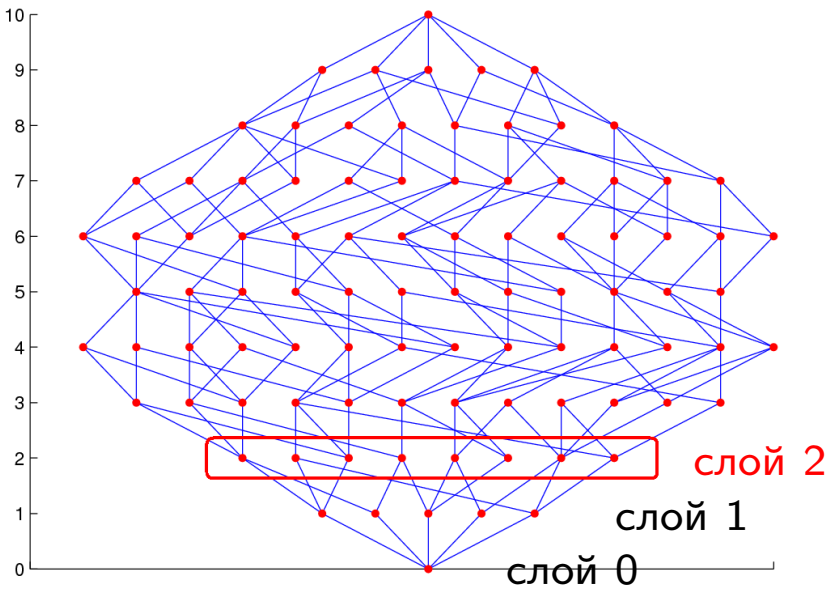




	слой 0	слой 1				
x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример. Семейство линейных алгоритмов классификации





	слой 0	слой 1					слой 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

Характеристики расслоения и связности алгоритма $a \in A$

Определение

Верхняя связность $u(a)$ алгоритма a — это число всех рёбер, исходящих из вершины a:

$$u(a) = |X_a|, \quad X_a = \left\{ x_{ab} \in \mathbb{X} \;\middle|\; a \prec b \right\};$$

X_a называется *порождающим множеством* алгоритма a.

Определение

Неполноценность $q(a)$ алгоритма a — это мощность множества объектов, соответствующих всем рёбрам на путях, ведущих в a:

$$q(a) = |X'_a|, \quad X'_a = \left\{ x \in \mathbb{X} \;\middle|\; \exists b \in A\colon\, b < a,\; I(b,x) < I(a,x) \right\};$$

X'_a называется *запрещающим множеством* алгоритма a.

Верхняя оценка вероятности переобучения

Теорема (Воронцов, Решетняк, Ивахненко, 2010)

Для любого монотонного метода μ, любых $\mathbb{X}$, A и $\varepsilon \in (0,1)$

$$Q_\varepsilon(\mu,\mathbb{X}) \leqslant \sum_{a\in A} \frac{C^{\ell-u}_{L-u-q}}{C^\ell_L}\, \mathcal{H}^{\ell-u,\,m-q}_{L-u-q}\left(\tfrac{\ell}{L}\big(m-\varepsilon k\big)\right),$$

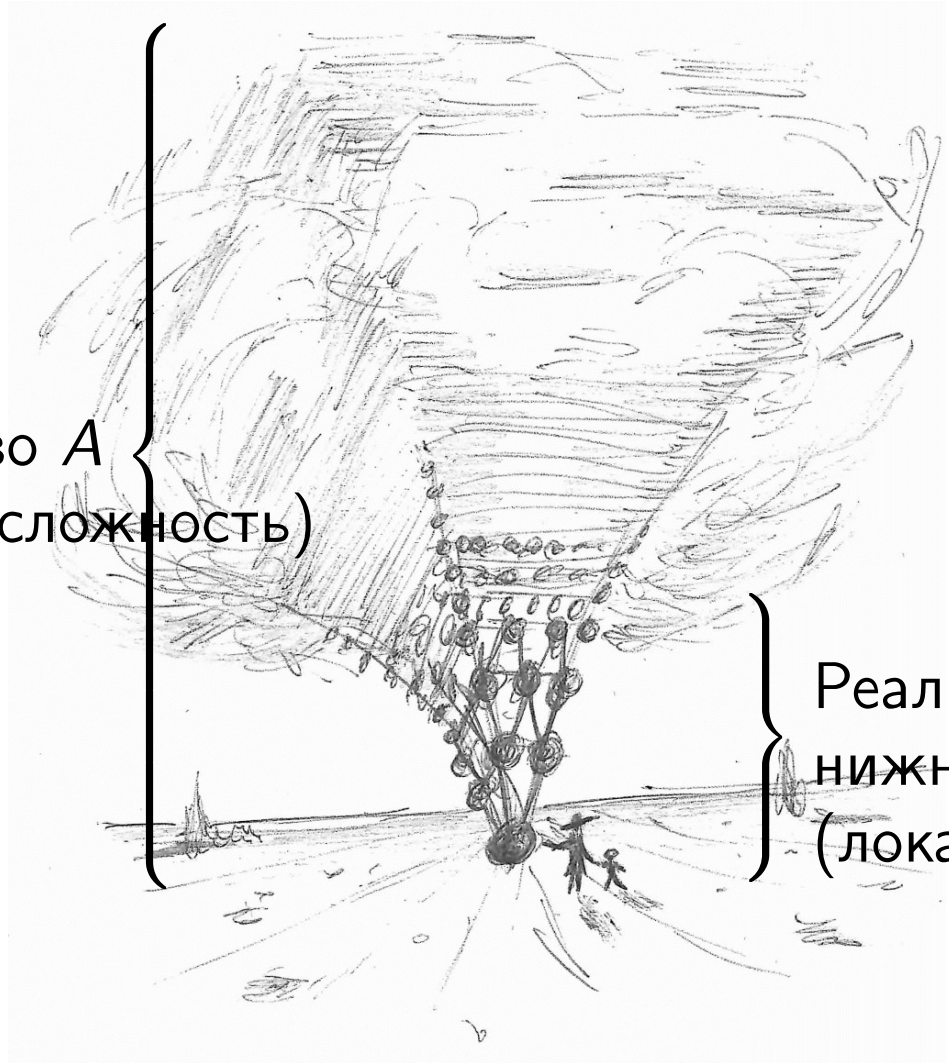
где $u = |X_a|$ — верхняя связность алгоритма a,
$q = |X'_a|$ — неполноценность алгоритма a,
$m = n(a,\mathbb{X})$ — число ошибок алгоритма a,

$$\mathcal{H}^{\ell,\,m}_L(z) = \sum_{s=0}^{\lfloor z\rfloor} \frac{C^s_m C^{\ell-s}_{L-m}}{C^\ell_L}, \quad z = 0,\ldots,\ell$$

— функция гипергеометрического распределения:

Следствие: $\mathsf{P}\big[\mu X = a\big] \leqslant C^{\ell-u}_{L-u-q} / C^\ell_L$.

Расслоение позволяет вычислять оценку по нижним слоям



Всё семейство A
(глобальная сложность)

Реально используемые
нижние слои A
(локальная сложность)

Завершённые и текущие исследования

- Верхние оценки CCV и Q_ε по графу расслоения-связности
- Модельные семейства A, для которых оценки CCV и Q_ε являются точными [Н.Животовский]
- Улучшение логических алгоритмов классификации [А.Ивахненко, П.Ботов, А.Фрей]
- Оценки переобучения симметричных семейств алгоритмов с помощью теоретико-группового метода орбит [А.Фрей]
- Оценивание переобучения путём покрытия семейства алгоритмов однослойными подсемействами [А.Фрей]
- Эффективное вычисление оценок переобучения с помощью случайных блужданий по графу [Е.Соколов]
- Точные оценки CCV для:
  метода ближайшего соседа [М.Иванов],
  монотонных классификаторов [А.Зухба, Г.Махина],
  одномерного порогового классификатора [Ш.Ишкина]

Ближайшие задачи (темы диссертаций)

- Оценивание характеристик графа расслоения-связности по случайной подвыборке.
- Обобщающая способность алгоритмов поиска закономерностей в символьных последовательностях.
- Обобщающая способность композиций одномерных пороговых классификаторов.

Литература

Воронцов К. В. Теория надёжности обучения по прецедентам. Курс лекций ВМК МГУ и МФТИ. 2011.
http://www.machinelearning.ru/wiki/images/d/d9/Voron-2011-tnop.pdf

Примеры прикладных задач матричных разложений

- Анализ данных жидкостной хроматографии

$$z(t,\lambda) = \sum_i X_i(t) Y_i(\lambda)$$

дано: $z(t,\lambda)$ — выход сканирующего УФ-детектора;
найти: $X_i(t)$ — хроматограмма i-го вещества,
$Y_i(\lambda)$ — спектр i-го вещества.

- Анализ данных ДНК-микрочипов

$$I(p,k) = \sum_g a_{pg} C_{gk}$$

дано: $I(p,k)$ — интенсивность свечения p-й пробы на k-м чипе;
найти: a_{pg} — коэффициент сродства p-й пробы g-му гену,
C_{gk} — концентрация g-го гена на k-м чипе.

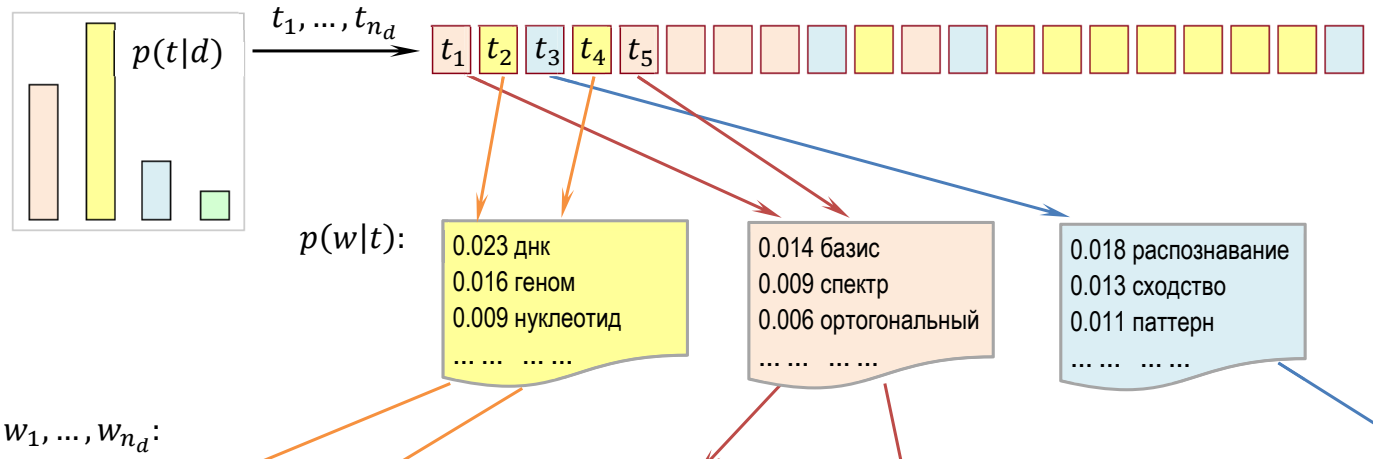
- Тематические модели коллекций текстовых документов

$$p(w|d) = \sum_t p(w|t) p(t|d)$$

дано: $p(w|d)$ — частоты слов w в документах d;
найти: $p(w|t)$ — распределения слов w в темах t,
$p(t|d)$ — распределения тем t в документах d.

Вероятностная модель порождения документа d

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Цели тематического моделирования (topic modeling)

- Тематический поиск документов и объектов по тексту любой длины или по любому объекту
- Категоризация, классификация, аннотирование, суммаризация текстовых документов

Типичные приложения:

- Поиск научной информации
- Поиск экспертов (expert search), рецензентов, проектов
- Выявление трендов и фронта исследований
- Анализ и агрегирование новостных потоков
- Рекомендательные сервисы (коллаборативная фильтрация)
- Рубрикация коллекций изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики

Вероятностный латентный семантический анализ PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Задача: найти максимум правдоподобия

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

$$\varphi_{wt} \geqslant 0; \quad \sum_{w\in W} \varphi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{d\in D} \theta_{td} = 1$$

Это задача поиска стохастического матричного разложения

$$\|F - \Phi\Theta\| \to \min_{\Phi,\Theta}$$

$F = \left(\hat{p}(w|d) = \frac{n_{dw}}{n_d}\right)_{W\times D}$ — матрица исходных данных;
$\Phi = \left(\varphi_{wt}\right)_{W\times T}$ — искомая матрица терминов тем $\varphi_{wt} = p(w|t)$;
$\Theta = \left(\theta_{td}\right)_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

EM-алгоритм

E-шаг: условные вероятности тем $p(t|d,w)$ для всех t, d, w вычисляются через φ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}.$$

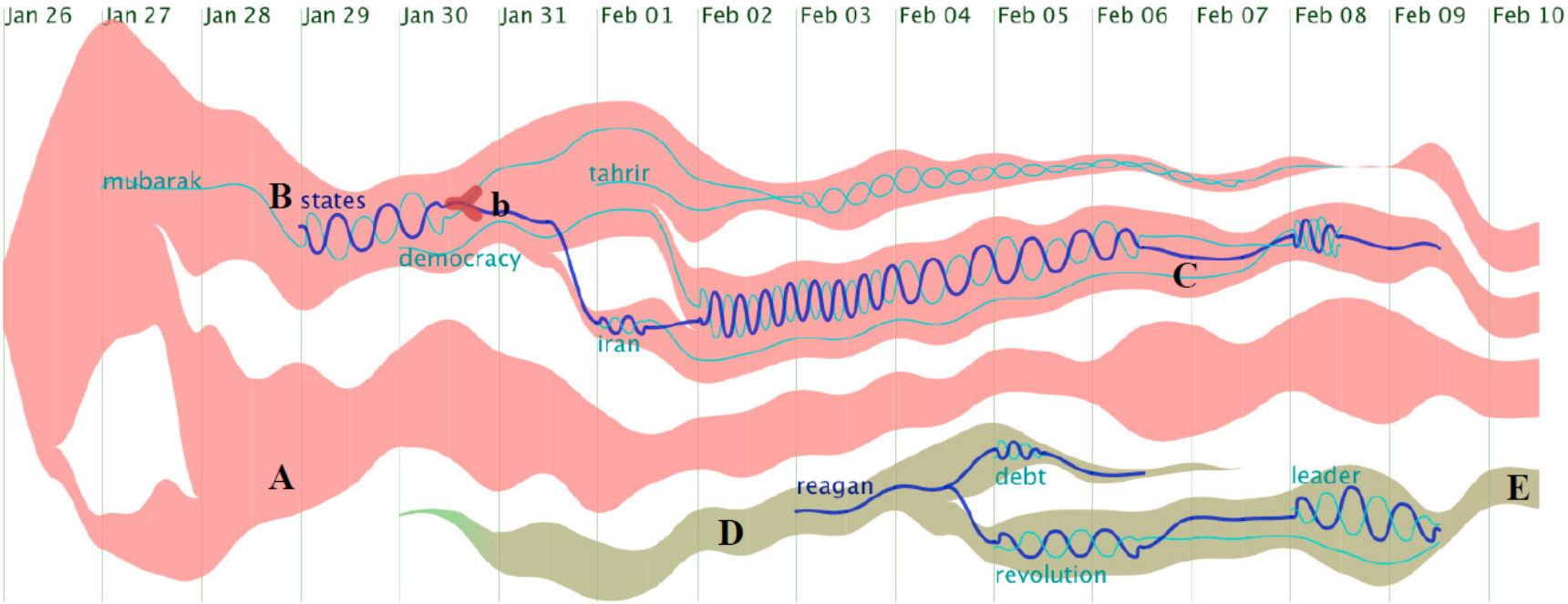
M-шаг: решение задачи максимизации правдоподобия выражается аналитически через частотные оценки условных вероятностей, если положить $n_{dwt} = n_{dw}p(t|d,w)$:

$$\varphi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d \in D} n_{dwt}, \qquad n_t = \sum_{w \in W} n_{wt};$$

$$\theta_{td} = \frac{n_{dt}}{n_d}, \qquad n_{dt} = \sum_{w \in d} n_{dwt}, \qquad n_d = \sum_{t \in T} n_{dt}.$$

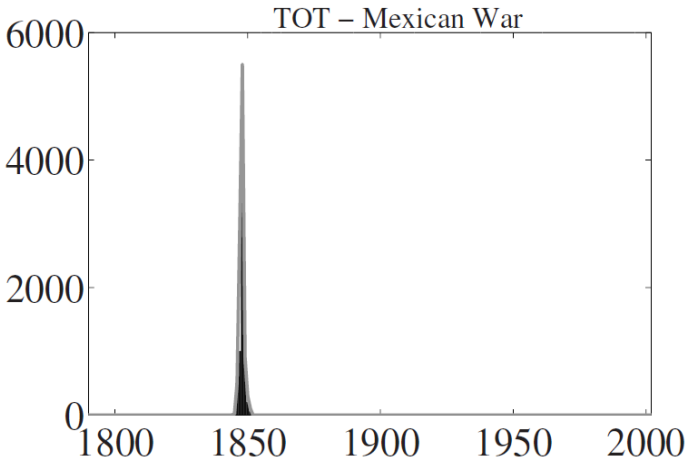
EM-алгоритм — это чередование E и M шагов до сходимости.

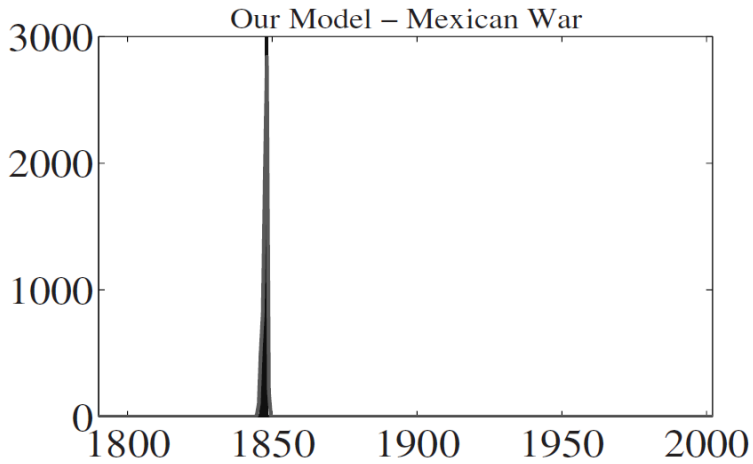
Пример. Динамическая тематическая модель



Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai J. Gao, Xin Tong, Huamin Qu TextFlow: Towards Better Understanding of Evolving Topics in Text // IEEE Transactions On Visualization And Computer Graphics, Vol. 17, No. 12, December 2011.

Пример. Совмещение динамической и n-граммной модели



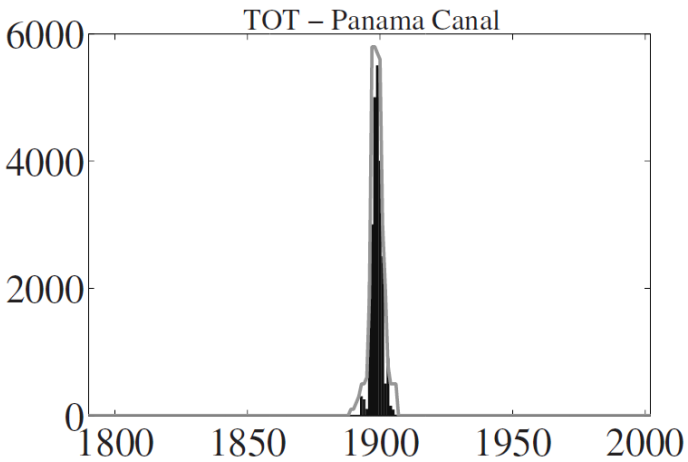


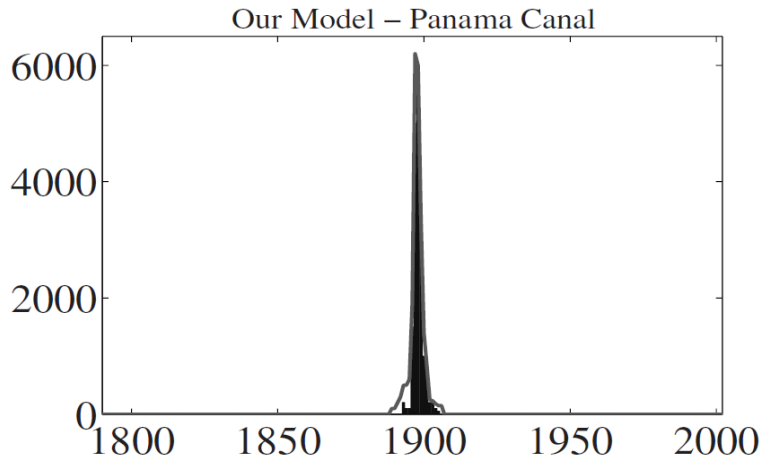
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Пример. Совмещение динамической и n-граммной модели



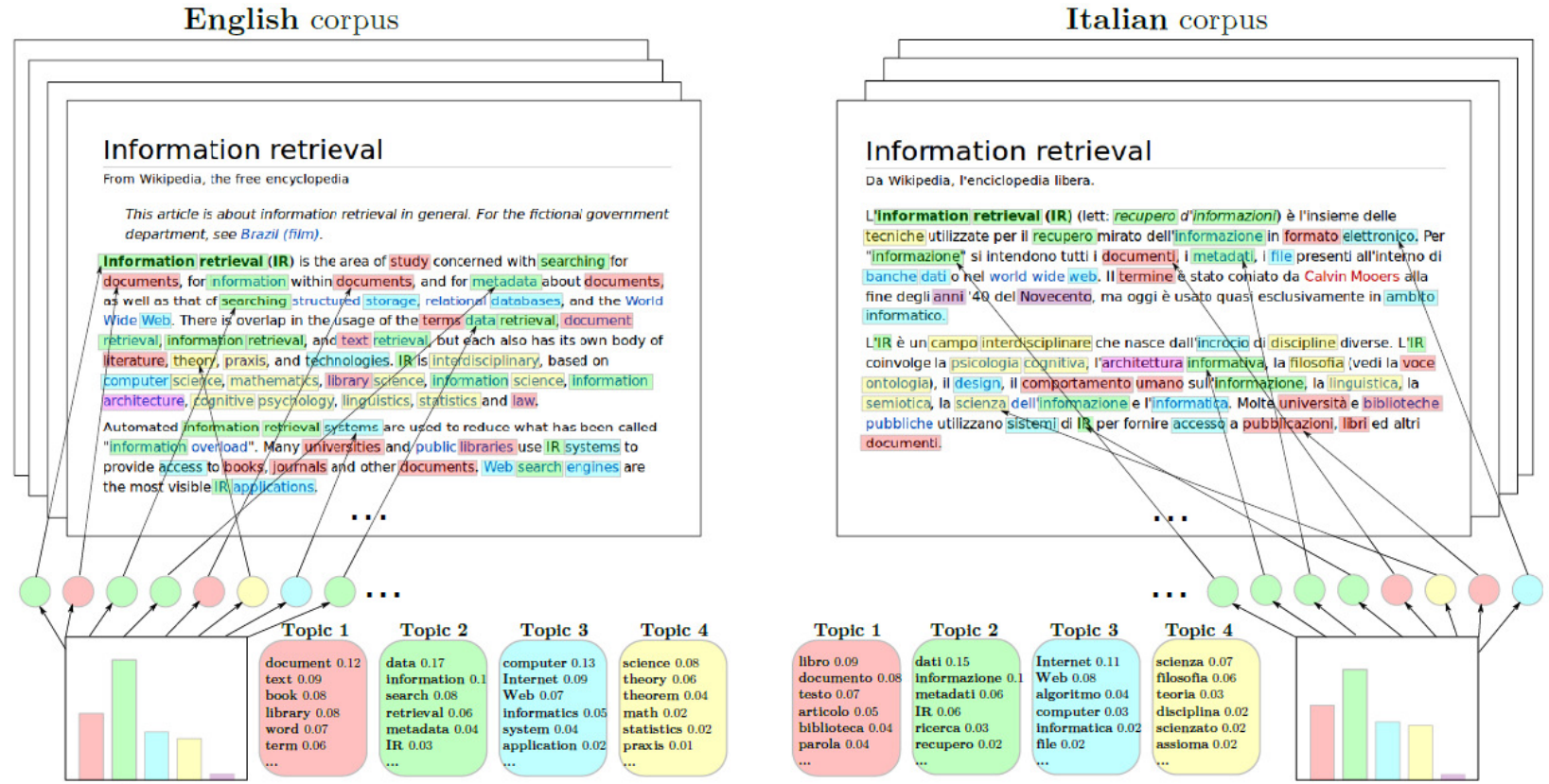


1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Пример. Многоязычные модели



I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications // NIPS, 7–8 December 2012. — Pp. 1–11.

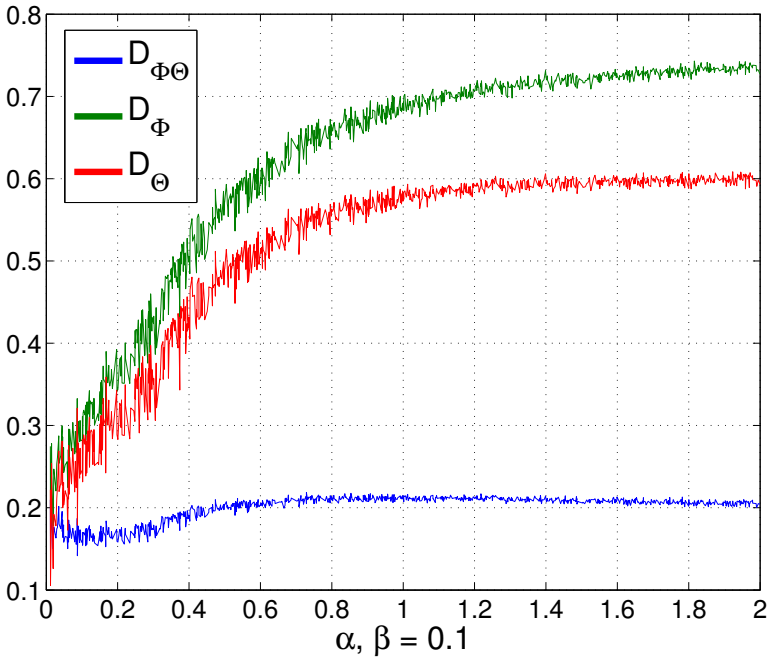
Завершённые и текущие исследования

- Аддитивная регуляризация тематических моделей (ARTM) — простая и мощная новая теория [К.Воронцов, 2013]
- Стартовал проект BigARTM — параллельная реализация общей модели ARTM на C++ [А.Фрей и др., 2014]
- Робастные алгоритмы, умеющие игнорировать нетематические слова [А.Потапенко,2012]
- Статистические критерии для проверки гипотезы условной независимости [В.Целых, 2013]
- Полностью автоматическое извлечение терминов для n-грамных моделей [С.Царьков, 2013]
- Эксперименты, показывающие неустойчивость моделей PLSA и LDA [В.Глушаченков, 2013]

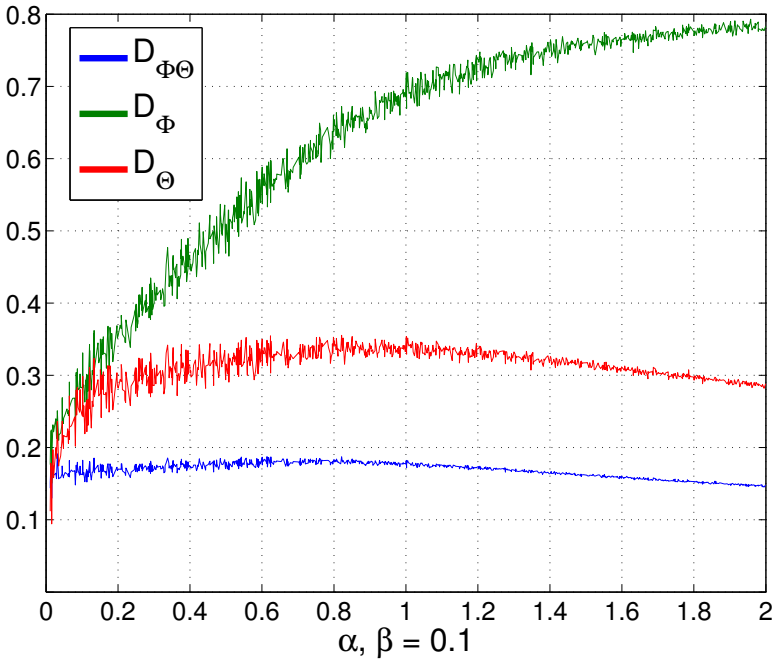
Неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ_0 [В.Глушаченков, 2013]:

PLSA

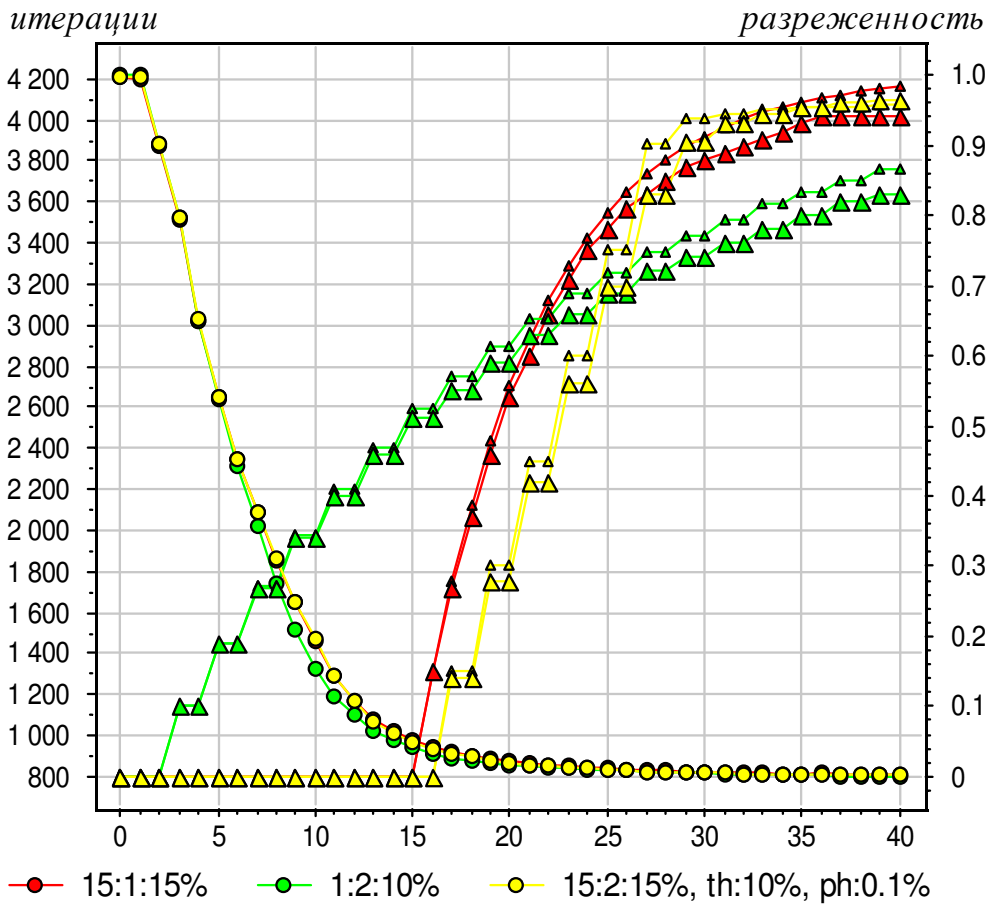


LDA



Разреживание распределений φ_{wt} и θ_{td}

Матрицы Φ и Θ можно разреживать на 95%, практически без потери точности модели [А.Потапенко, 2013]:



Аддитивная регуляризация тематических моделей (ARTM)

Задача максимизации регуляризованного правдоподобия:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

Решение: вместо обычных формул M-шага в EM-алгоритме

$$\varphi_{wt} \propto n_{wt}, \qquad \theta_{td} \propto n_{dt},$$

используются модифицированные

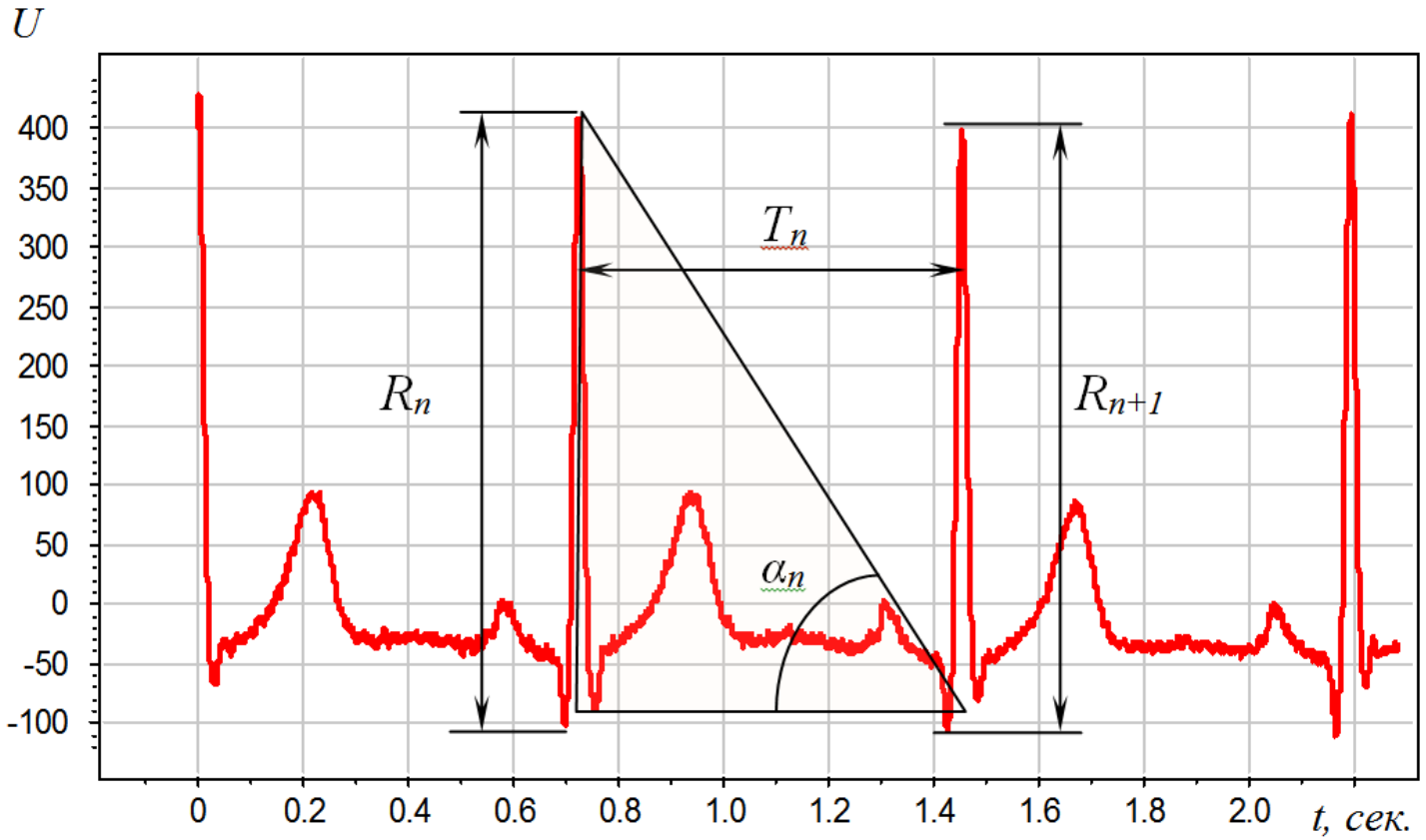
$$\varphi_{wt} \propto \left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right)_+, \qquad \theta_{td} \propto \left(n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+.$$

Достаточно продифференцировать R по параметрам

Ближайшие задачи (темы дипломных работ)

- Как улучшить интерпретируемость тем?
- Как определить число тем?
- Как сделать модель более устойчивой?
- Как учесть внешнюю информацию (время, авторы, ссылки, категории)?

Неклассический информационный анализ ЭКГ



Открытие профессора В.М.Успенского: ранняя диагностика многих заболеваний возможна по знакам приращений амплитуд $R_{n+1} - R_n$, интервалов $T_{n+1} - T_n$ и углов $\alpha_{n+1} - \alpha_n$.

Стадии анализа данных ЭКГ по В.М.Успенскому

1. вычисление амплитуд, интервалов и углов
   по кардиограмме длиной 600 кардиоциклов
2. вычисление *кодограммы* — 599-символьной строки
   в 6-буквенном алфавите
3. вычисление 216 признаков — частот триграмм
4. формирование эталонных выборок абсолютно здоровых
   и больных, для каждого заболевания
5. поиск *диагностических эталонов* — наборов триграмм,
   совместно встречающихся у больных данным заболеванием
6. обучение алгоритма классификации
7. статистическая оценка точности диагностики
   по контрольным выборкам или скользящему контролю
8. применение алгоритма классификации для диагностики

Кодограмма и частоты триграмм

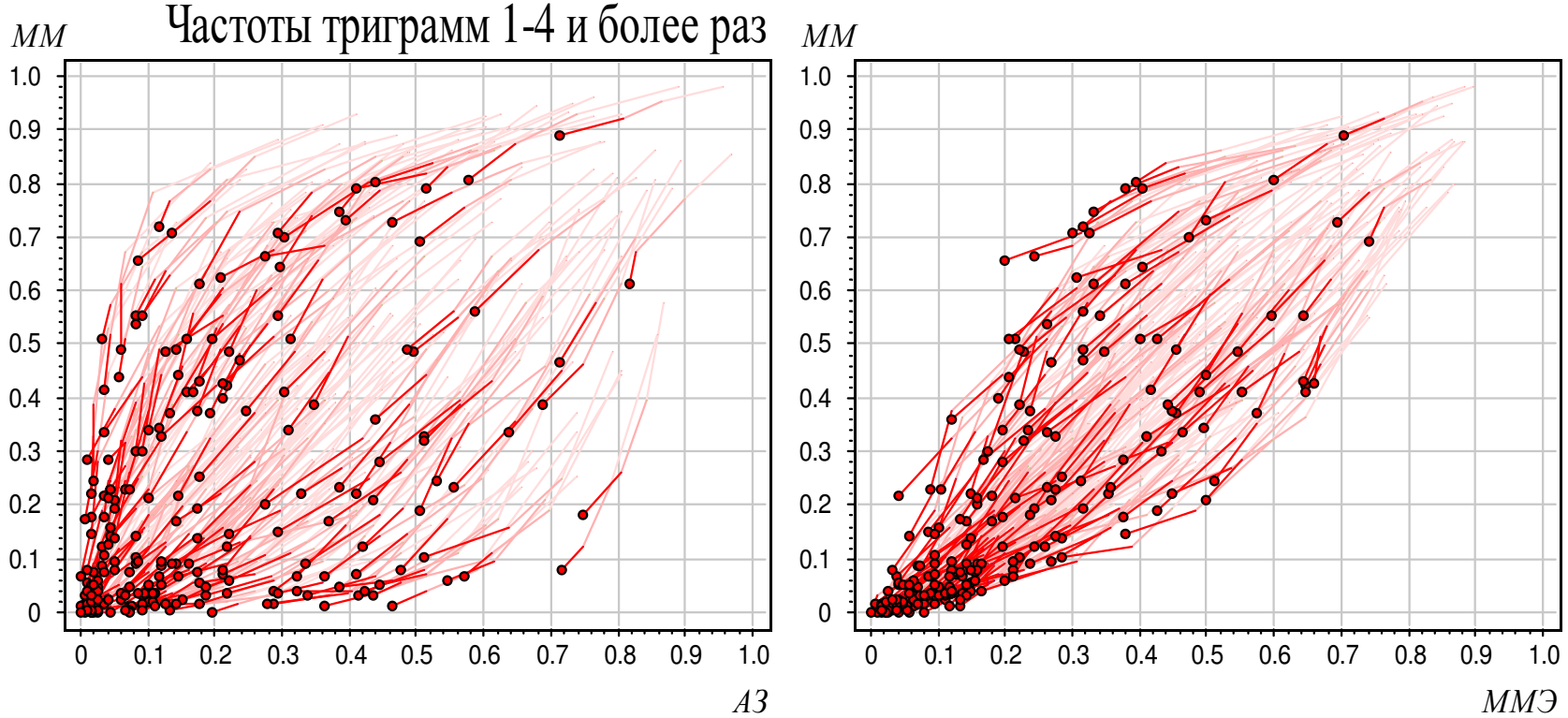
1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Диагностическая система «Скринфакс» (2-е поколение)

- более 10 лет эксплуатации
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 50 заболеваний
- из них более 20 имеют отобранные эталонные выборки

Поиск информативных признаков-триграмм

Слева: триграммы в осях «доля здоровых» — «доля больных».
Справа: триграммы в осях «доля больных» — «доля больных».



Вывод: информативные триграммы статистически значимы.

Линейный классификатор. Постановка задачи обучения

$(x^1, \dots x^n) \in \mathbb{R}$ — n числовых признаков объекта x;
Линейный алгоритм классификации:

$$a(x, \alpha) = \text{sign}\left(\sum_{j=1}^{n} \alpha_j x^j\right) = \text{sign}\langle \alpha, x \rangle.$$

Задача поиска вектора α решается путём минимизации сглаженного регуляризованного эмпирического риска:

$$Q(\alpha) = \sum_{i=1}^{\ell} \mathscr{L}\big(\langle \alpha, x_i \rangle y_i\big) + \frac{\tau}{2}\|\alpha\|^2 \to \min_{\alpha}.$$

Примеры функций потерь:

$$\mathscr{L}(M) = \begin{cases} \log(1 + e^{-M}) & \text{логистическая регрессия;} \\ (1 - M)_+ & \text{метод опорных векторов;} \\ e^{-M} & \text{композиция AdaBoost;} \end{cases}$$

Задача

Даны матрицы «объекты–признаки» по 5 болезням и здоровым, эталонные обучающие и контрольные выборки (всего 12 матриц).
Требуется построить алгоритм классификации и проверить его на контрольных данных.

Описание задачи — на странице

http://www.MachineLearning.ru/wiki/index.php?title=User:Vokov

раздел «Диагностика заболеваний по ЭКГ»

http://www.MachineLearning.ru/wiki/images/e/e3/Voron-2014-task-ekg.pdf

http://www.MachineLearning.ru/wiki/images/3/37/Voron-2014-task-ekg-data.rar

Что ещё есть на свете?

- Системы с алгоритмами машинного обучения:
  Python, scikit-learn — scikit-learn.org
  RapidMiner — rapidminer.com
  WEKA — www.cs.waikato.ac.nz/ml/weka
- Чемпионат мира по анализу данных:
  kaggle.com
- Репозиторий реальных задач UCI:
  archive.ics.uci.edu/ml
- Полигон алгоритмов классификации:
  Poligon.MachineLearning.ru
- Вики-ресурс на русском языке:
  www.MachineLearning.ru
  я там: «Участник:Vokov»

контакты:

Воронцов Константин Вячеславович
voron@forecsys.ru
www.MachineLearning.ru/wiki, «Участник:Vokov»

Стрижов Вадим Викторович
strijov@forecsys.ru
http://www.strijov.com
www.MachineLearning.ru/wiki, «Участник:Strijov»