«Современные проблемы прикладной математики»

О некоторых задачах интеллектуального анализа данных

Воронцов Константин Вячеславович

кафедра ММП
(Математические методы прогнозирования)

ВМиК МГУ, 28 апреля 2008

Почему наш анализ данных «интеллектуальный»?

- Алгоритмы обучаются по данным (прецедентам)
- Наш собственный интеллект тоже задействован ;)
- Попытка адекватного русского перевода термина «Data Mining»

Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [*Григорий Пятецкий-Шапиро*].

Задача обучения по прецедентам

X — *объекты* (описания наблюдений, ситуаций, случаев);
Y — *ответы* (значения откликов, прогнозов, решений).

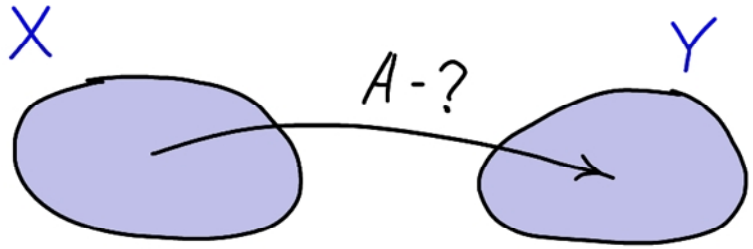
$x \equiv \left(f_1(x),...,f_n(x)\right)$ — данных об объекте $x \in X$,

где $f_j : X \to D_j$ — *признаки.*

$X^m = \left\{(x_1, y_1),...,(x_m, y_m)\right\}$ — выборка прецедентов, $y_i = y^*(x_i)$,

$y^* : X \to Y$ — неизвестная *целевая* зависимость.

Требуется построить
алгоритм $A : X \to Y$,
приближающий $y^*(x)$.



Алгоритм *обучается* (learns) по *обучающей* (training) выборке:
$\mu : X^m \mapsto A$ — метод обучения.

Типы задач обучения по прецедентам

Классификация $Y = \{0, 1\}$

- Медицинская диагностика
- Кредитный скоринг (credit scoring)
- Предсказание ухода клиентов (churn prediction)
- Фильтрация спама (spam filtering)
- Каталогизация текстов, новостей
- Распознавание плагиата
- Обнаружение мошенничества (fraud detection)
- Обнаружение биржевых интервенций

Регрессия, прогнозирование $Y = R$

- Предсказание цен акций, электроэнергии, и т.д.
- Прогнозирование объёмов продаж (sales forecast)
- Предсказание индивидуальных предпочтений (collaborative filtering)
- Моделирование и проектирование технических устройств

Тема для реферата

Обзор методов классификации, применяемых для предсказания ухода клиентов

Ключевые слова:
churn prediction

- Логистическая регрессия, решающие деревья, нейронные сети… что ещё? что лучше и почему?
- Для чего и как используются оценки вероятности ухода?
- Как планируются маркетинговые акции на основе сделанных предсказаний?

Методы классификации

- Линейный дискриминант Фишера (LDF)
- Метод k ближайших соседей (kNN)
- Метод парзеновского окна (Parzen window)
- EM-алгоритм
- Метод радиальных базисных функций (RBF)
- Метод потенциальных функций
- Метод опорных векторов (SVM)
- Метод релевантных векторов (RVM)
- Логистическая регрессия (LR)
- Нейронные сети (BackPropagation и др.)
- Решающие деревья (ID3, C4.5, CART, и др)
- Решающие списки (SCM)
- Индукция правил (ТЕМП, КОРА, IREP, RIPPER, SLIPPER)
- Алгоритма вычисления оценок (ABO)
- Бустинг, бэггинг, метод случайных подпространств
- Алгебраический подход Ю.И.Журавлёва

Индукция правил (rule induction)

$r : X \to \{0,1\}$ — *правило* (логическая закономерность);
$r(x) = 1 \Leftrightarrow$ r «покрывает» объект x.

Требования к закономерностям:

- *Интерпретируемость*:
  смысл правила $r(x)$ должен быть понятен эксперту.

- *Информативность*:
  правило $r(x)$ должно покрывать много объектов класса y
  и мало объектов всех остальных классов.

- *Взаимодополняемость*:
  композиция правил вместе должна образовать
  алгоритм классификации.

Интерпретируемость правил

Смысл правила $r(x)$ должен быть понятен эксперту:

$$r(x) = \mathop{И}_{j\in\Omega}\left[a_j \le f_j(x) \le b_j\right] \text{ — } \textit{конъюнктивное правило},$$

где $\Omega \subseteq \{1,...,n\}$ — набор признаков, $|\Omega| \lesssim 7$, a_j, b_j — пороги.

Пример 1:

ЕСЛИ Пол = мужской **И**
Семья = холост **И**
28 ≤ Возраст ≤ 34 **И**
РабочийТелефон = нет **ТО** **не выдавать кредит**

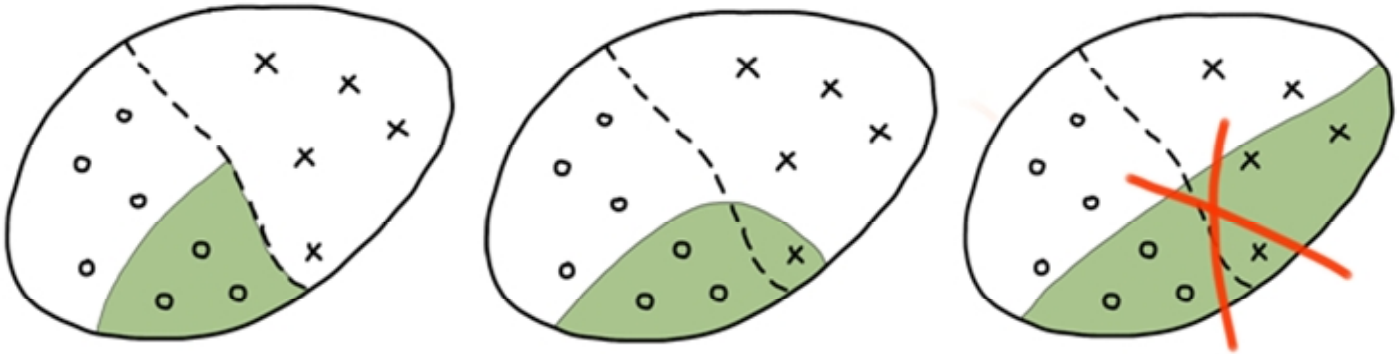
Пример 2:

ЕСЛИ Возраст > 60 **И**
РанееПеренёсИнфаркт = да **И**
АртериальноеДавление > 140 **ТО** **отложить операцию**

Информативность правил

$p_y(r)$ — число положительных примеров (покрытых класса y);

$n_y(r)$ — число отрицательных примеров (других классов);



Проблема: как определить функционал информативности?
Какую взять «свёртку» двух критериев

$$\begin{cases} p_y(r) \to \max \\ n_y(r) \to \min \end{cases}$$

Взаимодополняемость правил

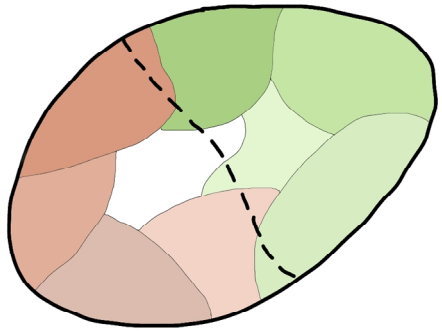
Задача классификации на два класса: $Y = \{0, 1\}$.

Набор правил (rule set) $\left\{ r_{y,t}(x) \,\middle|\, y \in Y,\ t = 1, \ldots, T_y \right\}$.

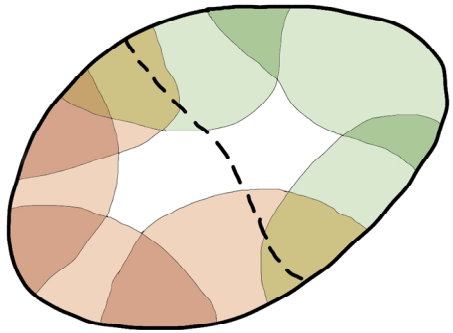
Алгоритм классификации — взвешенное голосование правил:

$$A(x) = \left[\sum_{t=1}^{T_1} w_{1,t} r_{1,t}(x) > \sum_{t=1}^{T_0} w_{0,t} r_{0,t}(x) \right],$$

где $w_{y,t}$ — неотрицательный вес правила $r_{y,t}(x)$ класса y.



Покрытие (решающие деревья и списки)



Голосование

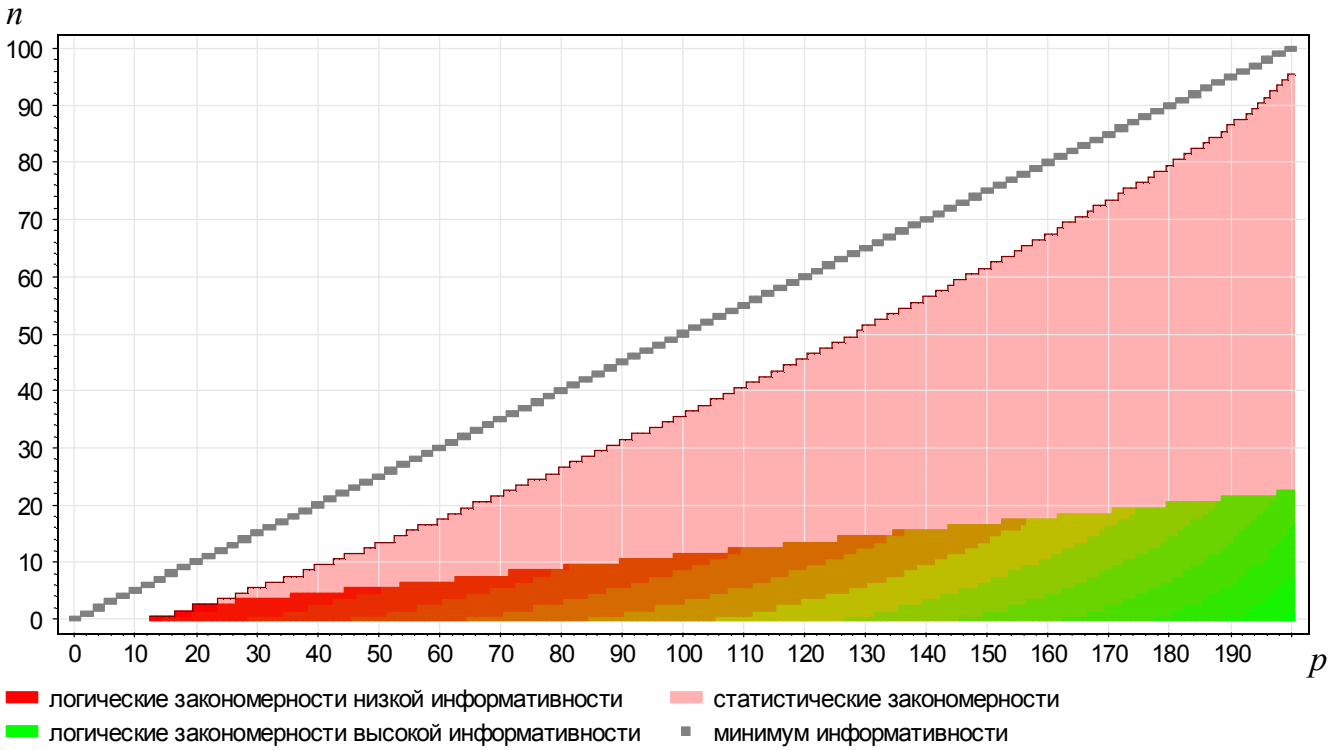
Трудность: задача о минимальном покрытии NP-трудная

Критерии информативности правил

- *Эвристический*: $\frac{p_y(r)}{m} > \delta$; $\frac{n_y(r)}{p_y(r)} < \varepsilon$
- *Статистический*:
  гипотеза H_0: $r(x)$ и $y^*(x)$ — независимые случайные величины;
  Тесты: χ^2, ω^2, FET (точный тест Фишера).
- *Энтропийный*: выигрыш информации (предельный случай FET).
- … уже придумано более 20 критериев [обзор Fürnkranz, 2005]
- *Мета-обучаемый* критерий (…через два слайда)

Эвристический и статистический критерии

(*p*,*n*)-*space* *при* *m* = 300, положительных *P* = 200, отрицательных *N* = 100



Вывод:

Неслучайность ещё не означает *закономерность*

Методы поиска покрывающего набора информативных конъюнкций

Все применяемые методы — более или менее удачные эвристические способы сокращения полного перебора.

- (Стохастический) Локальный поиск
- Метод ветвей и границ
- Генетические алгоритмы
- Случайный поиск с адаптацией

Все эти алгоритмы в обобщённо-упрощённом виде:

повторять
    взять следующий предикат $r(x)$;
    оценить его информативность;
    если он хороший **то**
        добавить его в список лучших закономерностей;
пока не выполнится критерий останова;

Мета-обучение (meta-learning)

Идея 1: Разбить выборку X^m на обучение X^l и контроль X^k

$p_y(r)$, $n_y(r)$ — на обучении;

$p'_y(r)$, $n'_y(r)$ — на контроле;

$Acc(r) = \dfrac{n'_y(r)}{p'_y(r) + n'_y(r)}$ — *точность* правила на контроле.

Идея 2: Искать регрессионную зависимость

$Acc(r) = f\left(p_y(r), n_y(r), \ldots\right)$;

обучающие *мета-объекты* — $\langle$задача, разбиение, правило$\rangle$.

Эксперименты [Fürnkranz, 2003-2007]:

140 000 мета-объектов;

30 реальных задач классификации из репозитория UCI;

f — нейросеть, либо линейная регрессия.

Результат: Мета-обучаемый критерий работает лучше!

Тема для реферата

Применение мета-обучения для выбора оптимальных эвристик в методах индукции правил

Ключевые слова: meta-learning, rule induction

Что ещё можно мета-обучать?

- параметры поисковых процедур: ширина поиска, количество поколений, критерии останова, темп адаптации, и т.п.;
- способ построения покрытия (используется ли жадный алгоритм или перевзвешивание объектов? Если второе, то функцию весов объектов логично подбирать мета-обучением. Есть ли работы в этом направлении?)

Fürnkranz J., Flach P. A. **Roc 'n' rule learning — towards a better understanding of covering algorithms** // *Machine Learning*. — 2005. — Vol. 58, no. 1. — Pp. 39–77.
http://dblp.uni-trier.de/db/journals/ml/ml58.html#FurnkranzF05

Janssen F., Fürnkranz J. **Meta-learning rule learning heuristics** // LWA / Ed. by A. Hinneburg. — Martin-Luther-University Halle-Wittenberg, 2007. — Pp. 167–174.
http://dblp.uni-trier.de/db/conf/lwa/lwa2007.html#JanssenF07

Проблема сверх-больших данных

«Сверх-большая» выборка — это $m \gtrsim 10^5$

Идея 1:

Sampling — случайная подвыборка $X^n \subset X^m$, $n \ll m$

Идея 2:

1. Понизить порог информативности;
2. Нагенерировать избыточно много правил;
3. Проверить их на $X^m \setminus X^n$ и неудачные отбросить;

Результат:

1. Можно вывести формулу: насколько надо понизить пороги, чтобы почти не потерять хорошие правила.
2. Находит почти всё, но может потребовать большого перебора.

Toivonen H. **Sampling large databases for association rules** // In Proc. 1996 Int. Conf. Very Large Data Bases / Ed. by T. M. Vijayaraman, A. P. Buchmann, C. Mohan, N. L. Sarda. — Morgan Kaufman, 1996. — Pp. 134–145.
citeseer.ist.psu.edu/toivonen96sampling.html

Тема для реферата

Методы индукции правил на сверхбольших выборках

Ключевые слова: sampling, rule induction

- Как работать со сверх-большими выборками — есть ли ещё идеи кроме самплинга?
- Делают ли самплинг по двум, трём и т.д. подвыборкам?
- Тойвонен применял самплинг для поиска ассоциативных правил. Это очень похоже но конъюнктивные правила, но немного не то. Найдите работы, посвящённые самплингу именно для правил.

Toivonen H. **Sampling large databases for association rules** // In Proc. 1996 Int. Conf. Very Large Data Bases / Ed. by T. M. Vijayaraman, A. P. Buchmann, C. Mohan, N. L. Sarda. — Morgan Kaufman, 1996. — Pp. 134–145.
citeseer.ist.psu.edu/toivonen96sampling.html

Оценивание вероятностей и рисков

Задача кредитного скоринга:

- ОжидаемаяПотеря = СуммаКредита × ВероятностьДефолта

Задача предсказания ухода клиента:

- ОжидаемаяПотеря = ДоходностьКлиента × ВероятностьУхода

Идея 1:

Вероятность оценивается по контрольной выборке

Идея 2:

Оценка апостериорной вероятности $\hat{P}(y \mid x)$:

$$\hat{P}(y \mid x) = \frac{\textit{число точек } x_i : (y_i = y) \textit{ в окрестности } x}{\textit{число точек } x_i \textit{ в окрестности } x}.$$

Как строить окрестность?

Идея 3:

- берём все правила: $r_{y,t}(x) = 1$;
- для каждого из этих правил берём K ближайших к x точек x_i, оценивая близость по набору признаков Ω;

Тема для реферата

Методы оценивания апостериорных вероятностей в логических алгоритмах классификации

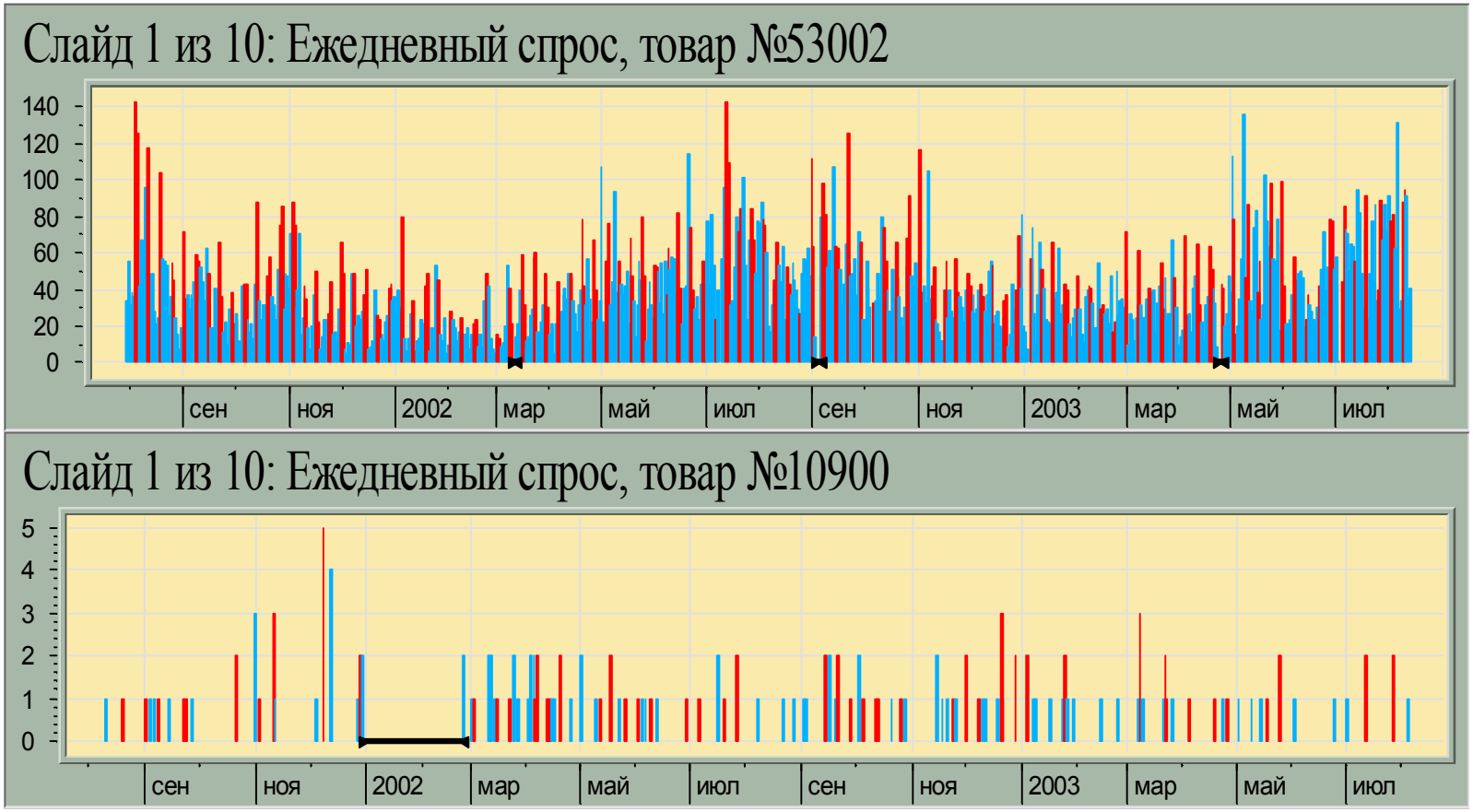
Ключевые слова:
rule induction, probabilistic output, probabilistic calibration

- Применяется ли калибровка Платта для логических правил?
- Какие есть методы преобразования бинарных логических правил в вероятностные (нечёткие) правила?
- Какие методы используются чаще: параметрические (логит-анализ, пробит-анализ) или непараметрические (сглаживание)?

J. Platt. **Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods** // Advances in Large Margin Classiers, MIT Press, 1999.
http://citeseer.ist.psu.edu/platt99probabilistic.html

Прогнозирование временных рядов продаж

- 40 000 товаров, 200 магазинов — 10^7 рядов
- Не существует хорошей модели ряда
- **Цель** — не точность прогнозов, а минимизация потерь



Неклассическая функция потерь (несимметричная, неквадратичная)

$\sum_{t=1}^{T} F\left(A(t) - y^*(t)\right) \to \min_A$, где *Потери, руб* $= F($*ошибка прогноза*$)$.

Классика: $F(\varepsilon) = \varepsilon^2$, но у нас $F(\varepsilon)$ не такая:

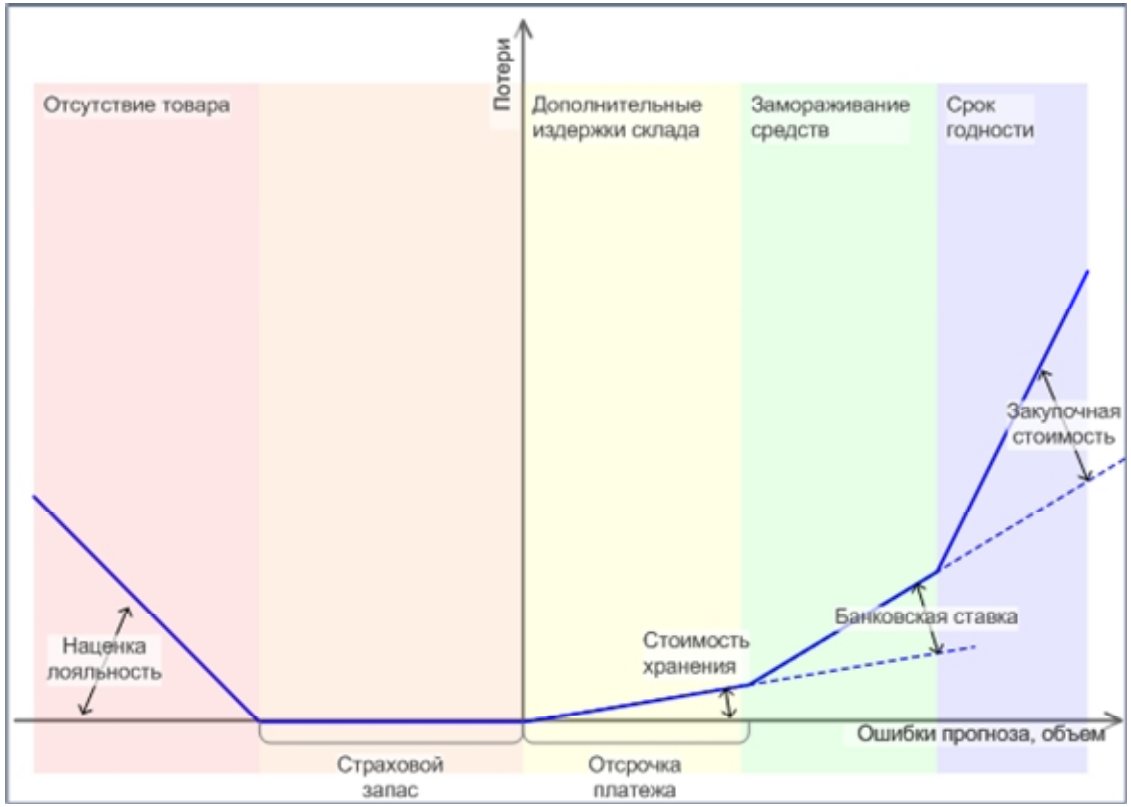


Схема решения

Идея: прогнозировать не значение ряда $y(t)$, а плотность распределения $p(y \mid t)$ (density forecast).
Для её реализации необходимо, чтобы ряд был стационарным.

- простыми быстрыми методами снять сезонность, тренды, праздники, выбросы, и т.п., получить стационарный ряд;
- спрогнозировать плотность распределения $p(y \mid t)$;
- найти оптимальное прогнозное значение $A(t)$ по критерию минимума ожидаемых потерь:

$$A(t) = \min_a \int F(a - y) p(y \mid t) dy$$

- к найденному прогнозу добавить ранее вычтенные факторы сезонности, тренда, праздников, выбросов, и т.п.

Тема для реферата

Обзор методов, применяемых для прогнозирования объёмов продаж

Ключевые слова:
sales forecast, density forecast, asymmetric loss

- Нейросеть и векторную авторегрессию не предлагать!
- Обзор методов, дающих прогнозы в виде плотности распределения возможных значений (density forecast).
- Как учитываются взаимозависимости товаров при прогнозировании?

Yong Bao, Tae-Hwy Lee, Burak Saltoglu ***Comparing Density Forecast Models***, 2006.

Stephen G. Hall, James Mitchell ***Density Forecast Combination.*** 2004.

Анализ клиентских сред, Коллаборативная фильтрация

U — множество клиентов (пользователей — users);
R — множество ресурсов (услуг, предметов, товаров — items)
$D = \{u_i, r_i, v_i\}_{i=1}^{m}$ — протокол пользования (выборка!),
v_i — объём пользования, возможны варианты:

- 0,1 — нет/да;
- рейтинг — целое число из {1,…,5} или {1,…,10};
- сумма покупки.

Пример: Задача конкурса Netflix (www.netflixprize.com):

- 480 189 клиентов, 17 770 фильмов, 10^8 рейтингов {1,…,5};
- Приз $\$10^6$ за увеличение точности прогнозов на 10%
- Функционал — root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{m'}\sum_{i=1}^{m'}\left(A(u_i', r_i') - v_i'\right)^2},$$

где $D' = \{u_i', r_i', v_i'\}_{i=1}^{m'}$ — контрольная выборка

Тема для реферата

Обзор методов коллаборативной фильтрации, применяемых лучшими участниками конкурса Netflix.

Ключевые слова:
collaborative filtering, personalization,
recommendation systems, web usage mining

- Какие методы не оправдали себя?
- Метод главных компонент работает, что ещё?
- Применяются ли байесовские модели?

Цели анализа клиентских сред

- Персонализация предложения
- Автоматическая каталогизация ресурсов (документов), в том числе персональная
- Поиск схожих ресурсов (документов)
- Поиск схожих клиентов (like minded people)

Основные подходы

- Анализ клиентов (user-based CF)
- Анализ ресурсов (item-based CF)
- Совмещение (ко-кластеризация и др.)
- Байесовские (латентные) модели

Байесовская модель посещений

Пусть X — множество тем.
Можно ли узнать по протоколу посещений, кто какой темой интересуется, и каким темам соответствует каждый ресурс?

$p(x \mid u)$ — скрытый тематический профиль клиента u;
$q(x \mid r)$ — скрытый тематический профиль ресурса r;

$$p(u,r) = \sum_t p(u)p(x \mid u)q(r \mid x) = \sum_t q(r)q(x \mid r)p(u \mid x)$$

По формуле Байеса:

$$q(r \mid x) = \frac{q(x \mid r)q(r)}{\sum_s q(x \mid s)q(s)}, \quad p(u \mid x) = \frac{p(x \mid u)p(u)}{\sum_v p(x \mid v)p(v)}$$

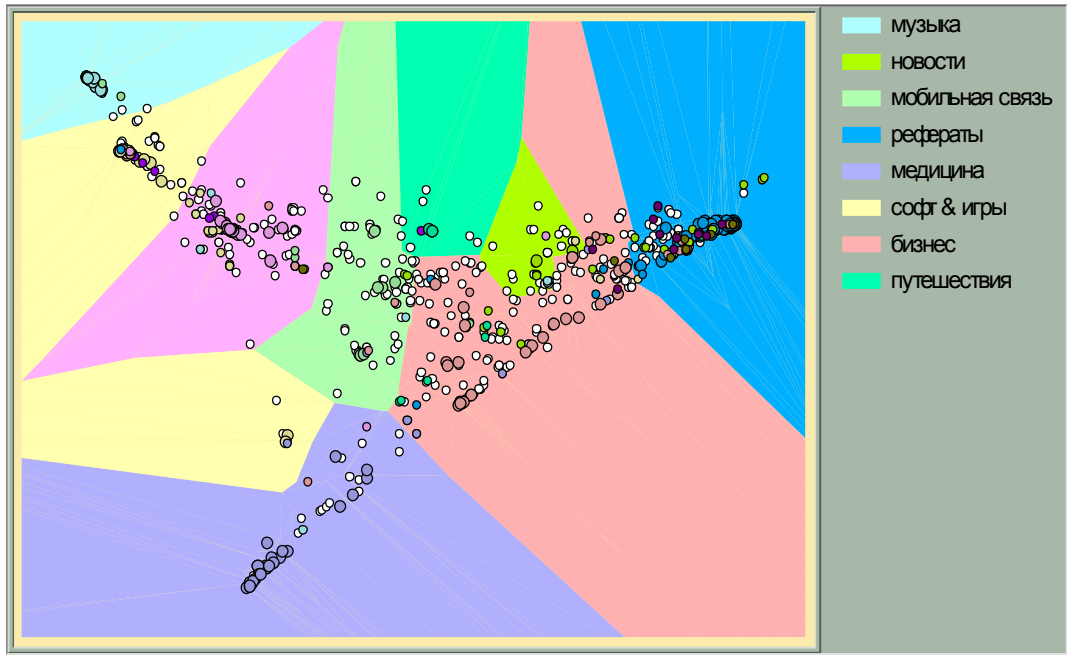
После этого $p(u,r)$ зависит только от скрытых профилей.
Задача максимизации правдоподобия:

$$\sum_{i=1}^{m} \ln p(u_i, r_i) \to \max_{p(x|u),\, q(x|r)}$$

Зачем нужны эти «профили»?

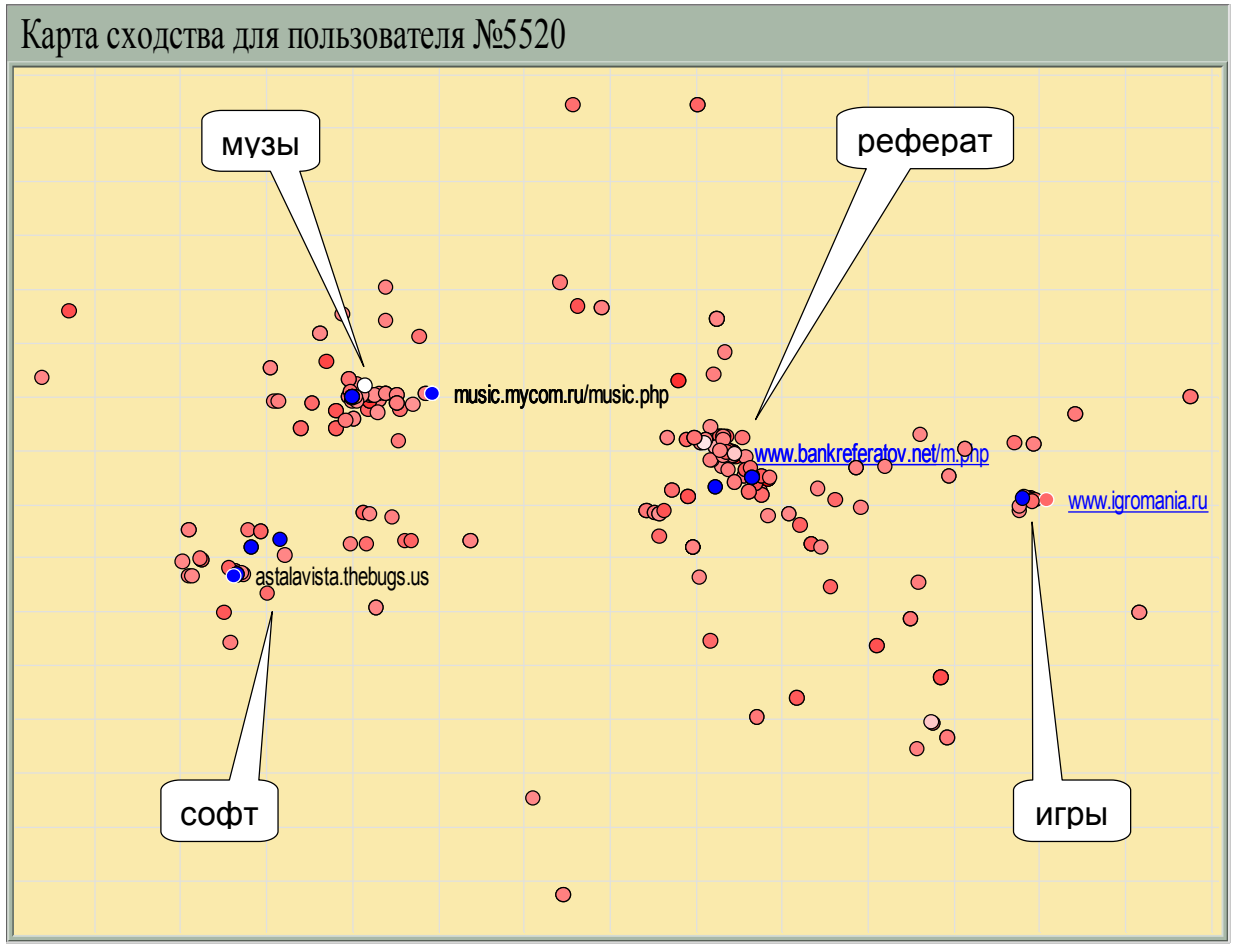
- Гораздо эффективнее сравниваются клиенты, ресурсы
- Меньше памяти для хранения данных
- Можно сравнивать клиента с ресурсом
- Тематические профили хорошо интерпретируемы
- Решается проблема «холодного старта»
- Возможно «частичное обучение», когда профили задаются лишь у некоторых ресурсов

Карта сходства ресурсов Интернет:



Зачем нужны эти «профили»?

Персональная карта сходства:
предложение ресурсов пользователю



Тема для реферата

Обзор методов коллаборативной фильтрации, использующих восстановление скрытой информации.

Ключевые слова:
generative model for CF, latent class models for CF

- Предупреждение: скрытая информация не обязательно называется «тематическими профилями»

Много ссылок на материалы по CF:
http://ict.ewi.tudelft.nl/~jun/CollaborativeFiltering.html

Анализ текста (Text mining)

Обнаружение заимствований (плагиата)
Пример: Система Антиплагиат, www.antiplagiat.ru

Темы для рефератов:

Обзор методов, применяемых для поиска заимствований.

Обзор методов, применяемых для оценивания сходства (релевантности) текстов.

Проблема:

- Как отличить плагиат от «законного» цитирования?
- Как свести эту задачу к обучению по прецедентам?

Эту презентацию и дополнительные ссылки можно найти на www.MachineLearning.ru

Страница
Участник:Vokov

Подстраница
Участник:Vokov/Некоторые задачи интеллектуального анализа данных (лекция)

Рефераты сдаются

- в бумажном виде на кафедру ММП
- в электронном виде мне, vokov@forecsys.ru