Сегментация транзакционных данных розничных клиентов банка

Козлинский Евгений

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем
Научный руководитель профессор РАН, д.ф.-м.н. К. В. Воронцов

13.06.19

Постановка задачи

Дано

- $d = w_{d,1}, \ldots, w_{d,N_d}$ - история транзакционных данных клиента
  $d \in D$ - клиент из набора D
  N_d - количество транзакций клиента d

Найти

- Тематическое векторное представление клиента
- Моменты изменения потребительского поведения

Критерии качества сегментации

P_k и WindowDiff

Цель исследования

Проблема

- Отсутствие размеченных данных
- Отсутствие четких критериев изменения потребительского поведения

Цель работы

- Получить векторное представление истории транзакций для проведения сегментации
- Разработать метод оценивания качества сегментации истории транзакций клиента
- Проверить гипотезу о незначительном изменении качества сегментации при переходе к тематическому представлению истории клиента.

Постановка задачи тематического моделирования

Дано: W - словарь mcc-кодов транзакций
D - коллекция историй транзакций пользователей d
Матрица $F = \{n_{dw}\}_{W \times D}$
n_{dw} - сумма покупок клиента d по коду w
T - множество тем потребительского поведения клиентов

Найти: Матрицы $\Phi = \{\phi_{wt}\}_{W \times T}, \quad \Theta = \{\theta_{td}\}_{T \times D}$
$\phi_{wt} = p(w|t)$ - вероятность кода w в теме t
$\theta_{td} = p(t|d)$ - вероятность темы t у клиента d

Гипотеза условной независимости: $p(w|d,t) = p(w|t)$
Из неё и формулы Байеса: $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$
Это задача матричного разложения: $F = \Phi\Theta$

Постановка задачи тематического моделирования

- У этой задачи $\exists$ бесконечно много решений вида $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$, где S - матрица ранга $|T|$
- Поэтому вводится регуляризация матриц Φ и Θ
  $PLSA: \quad R(\Phi,\Theta) = 0$
  $LDA: \quad R(\Phi,\Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$
  $ARTM: \quad R(\Phi,\Theta) = \sum_{i=1}^{N} \tau_i R_i(\Phi,\Theta)$
- **Задача оптимизации:**
  $\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$
  при условиях:
  $\sum_{w\in W} \phi_{wt} = 1; \quad \phi_{wt} \geq 0; \qquad \sum_{t\in T} \theta_{td} = 1; \quad \theta_{td} \geq 0$
- $p(t|w)$ - **тематика** mcc-кода

Описание алгоритма сегментации

Первый этап

- Выделение множества тем T потребительского поведения клиентов
- Представление транзакционных данных клиента последовательностью тематических векторов

Второй этап

Сегментирование последовательности векторов с помощью алгоритма Topic Tiling, определение числа сегментов

Обзор предыдущих работ

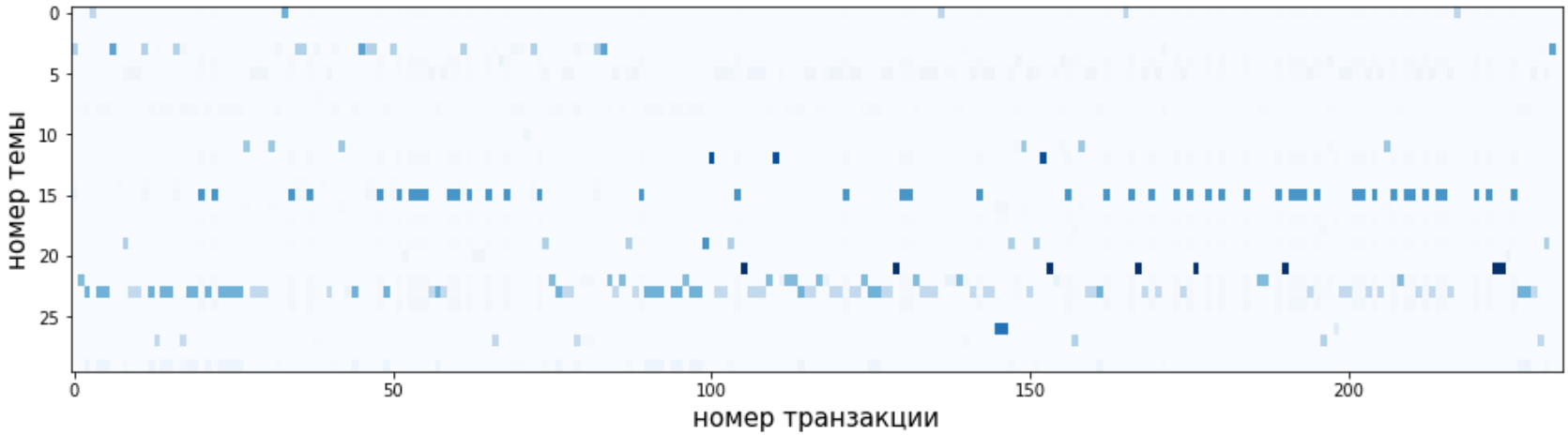
1. Ф.Никитин 2018 "Применение мультимодальных тематических моделей к анализу транзакционных данных":
2. Е.Смирнов 2018 "Тематическая сегментация диалогов контактного центра":
3. M.Riedl and C.Biemann 2012 "TopicTiling: A Text Segmentation Algorithm based on LDA"

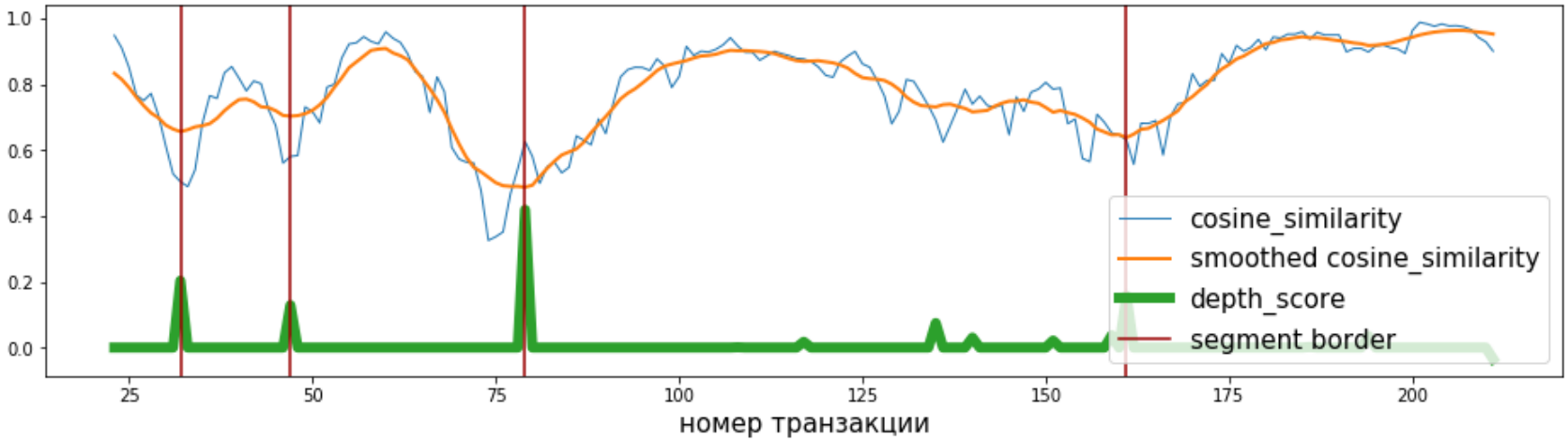
Сегментация с помощью Topic Tiling

Вход: последовательность транзакций (профиль), представленных в векторном виде.

- Проходим по профилю двумя скользящими окнами $(i - h_1, h_1]$ и $(h_1, i + h_1]$, для каждого i вычисляя cos-близость между средним левого и правого окна.
- Сглаживаем график cos-близости от i окном h_2.
- Для всех локальных минимумов на графике cos-близости считаем уверенность проведения сегмента в i: $depth\_score(i) = \frac{1}{2}(hl(i) - c_i + hr(i) - c_i)$, где c_i - значение косинусной близости в i, $hl(i)$ - ближайший к i локальный максимум слева, а $hr(i)$ - справа.
- $\mu = \hat{\mathbb{E}}_i(depth\_score(i))$, $\sigma = \hat{\mathbb{D}}_i(depth\_score(i))$, если $depth\_score(i) > \mu + \frac{\sigma}{2}$ проводим в i границу сегмента.

Сегментация с помощью Topic Tiling





а) Векторное представление одного профиля (сверху)

б) Стадии работы Аналога Topic Tiling (снизу)

Невязка между истинной и предсказанной сегментацией

$$Penalty(d) = \frac{1}{N_d - k} \sum_{i=1}^{N_d - k} [b_{s_{d,true}}(i) \neq b_{s_d}(i)]$$

С помощью P_k меры

$$b_{s_d}(i) = [w_{i,d} \in s_{q,d}][w_{(i+k),d} \in s_{q,d}],$$

где $s_{q,d}$ - сегмент сегментации s_d профиля d.

С помощью WindowDiff

$$\begin{cases} w_{d,i} \in s_{d,q} \\ w_{d,i+k} \in s_{d,t} \end{cases} \Longrightarrow b_{s_d}(i) = t - q$$

где $s_{d,q}$ и $s_{d,t}$ - сегменты сегментации s_d профиля d с порядковыми номерами q и t соответственно.

Данные и модели векторных представлений

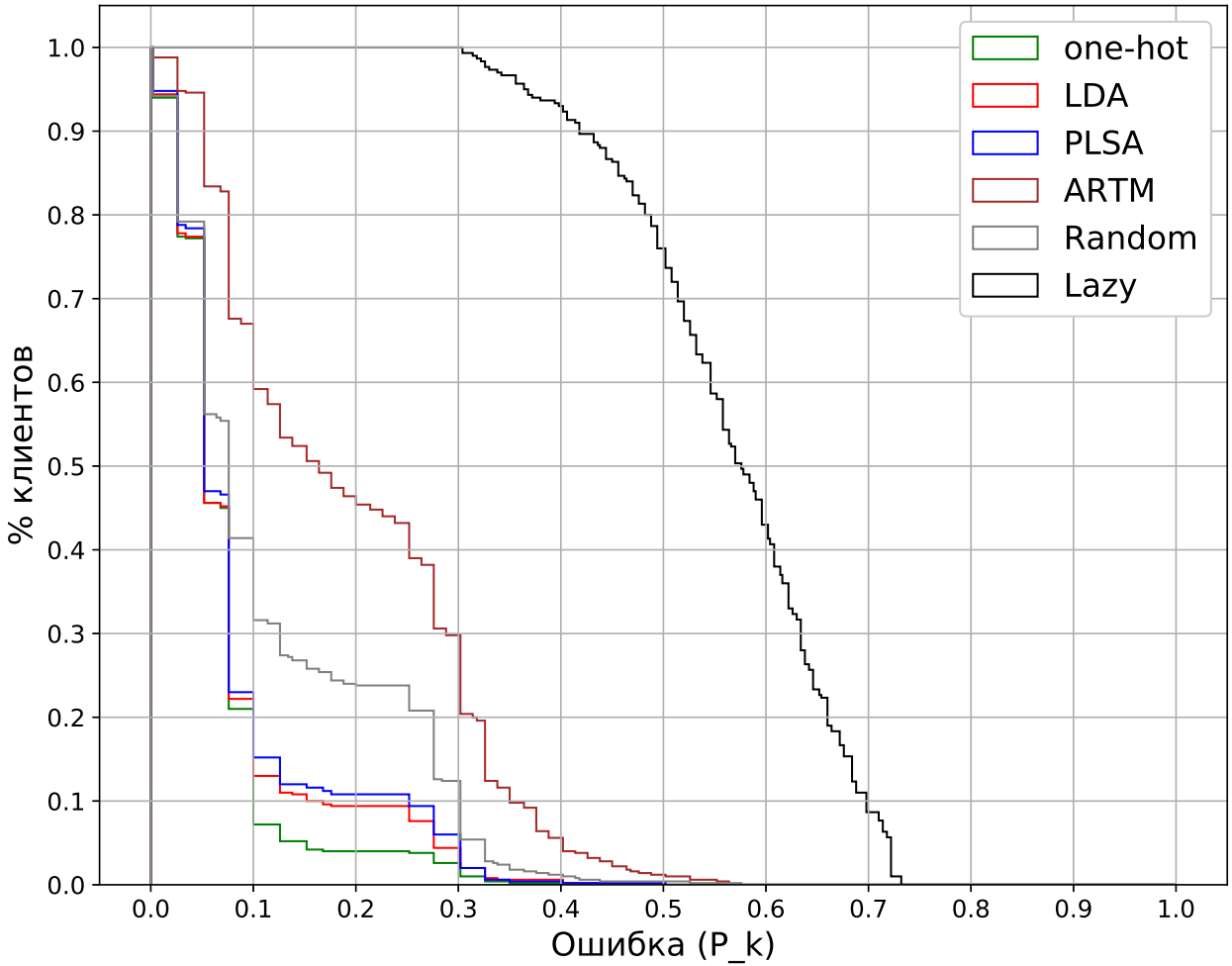
Данные: Транзакции $\sim$ 25000 пользователей за 3 года в виде таблицы с полями:

- *mcc_code* - mcc-код транзакции
- *amount_ru* - сумма транзакции в рублях
- *cardnumber* - идентификатор карты пользователя
- *trans_time* - дата и время транзакции

Модели векторных представлений:

- **PLSA** - PLSA на 30 темах
- **LDA** - LDA на 30 темах
- **ARTM** - ARTM на 30 темах, субъективно лучшие темы
- **one-hot** - тематика транзакции задана вектором с единицей на месте идентификатора mcc-кода
- **random** - тематика транзакции каждого mcc-кода задана случайным вектором
- **lazy** - модель не проводит границы сегментов

Эксперимент №1



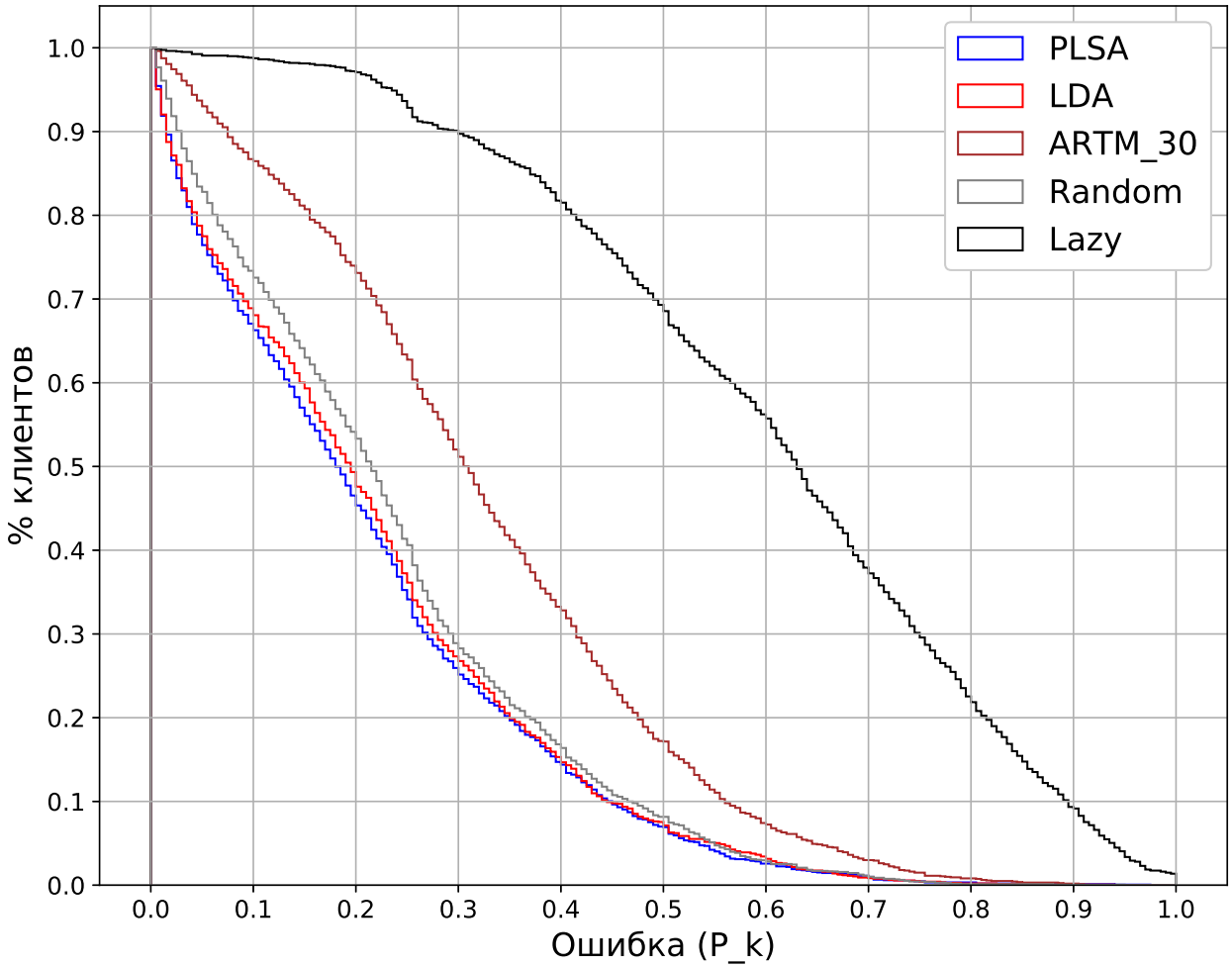
Точность проведения сегментов по сравнению с истинными на 500 искусственных профилях

Эксперимент №1

	P_k		WindowDiff	
Модель	mean	std	mean	std
lazy	0.566	0.103	0.566	0.103
one-hot	0.083	0.080	0.084	0.089
random	0.118	0.109	0.128	0.118
LDA	0.079	0.075	0.080	0.076
PLSA	0.083	0.080	0.084	0.081
ARTM	0.192	0.128	0.213	0.147

Точность проведения сегментов по сравнению с истинными на 500 искусственных профилях

Эксперимент №2



Результаты сравнения тематического сегментирования с one-hot на 2000 реальных профилях

Заключение

- Предложен способ представления истории пользователя с помощью тематик его транзакций для последующей сегментации
- Предложен метод оценивания качества модели сегментации транзакционных данных розничных клиентов
- Подтверждена гипотеза о незначительном изменении качества сегментации при переходе к векторным представлениям, для искусственных профилей