Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)

Факультет Управления и Прикладной математики
Кафедра Интеллектуальные Системы

Выпускная квалификационная работа бакалавра по направлению 03.03.01
«Прикладные математика и физика»

Сегментация транзакционных данных розничных клиентов банка

Студент 574 группы
Козлинский Е. М.

Научный руководитель
Воронцов К. В.

Долгопрудный
2019

Содержание

Глава 1

Введение

Работа посвящена построению тематической сегментации транзакционных данных для розничных клиентов банка на основе PLSA [1], LDA и ARTM [2] моделей с помощью вариации алгоритма Topic Tiling [3].

Ставится задача по нахождению моментов времени изменения потребительских привычек клиентов.

Поставленную задачу можно разбить на два этапа.

Этап первый: выделение набора тем и представление последовательностей транзакционных данных клиентов как последовательность тематик транзакций. Тематику транзакций можно получить из тематической модели. В работе [**?**](бакалаврская квалификационная работа Никитин) Была показана применимость подходов тематического моделирования к построению профиля розничных клиентов банка по транзакционным данным. В ней были построены PLSA, LDA и ARTM модели по транзакционным данным.

Этап второй: сегментирование последовательности векторов. В работе [**?**](Магистерский диплом Смирнов) решается схожая задача сегментации диалогов контактного центра. Ключевое отличие от нашей задачи - на втором этапе требуется выделение монотематических сегментов. В нашей работе это ограничение отсутствует.

Предлагается оценивать близость двух подпоследовательностей векторов с помощью косинусного расстояний их средних.

Для разрешения второго этапа используется аналог Topic Tiling.

Сложность заключается в отсутствии размеченных данных и отсутствии четких критериев, изменения потребительских привычек. Поэтому ставится два эксперимента. Первый определяет качество проведения сегментов аналогом Topic Tiling на искусственных профилях(границы сегментов в которых мы знаем). Второй эксперимент показывает насколько результат тематической сегментации реальных профилей клиентов похож на обычную сегментацию. В обоих экспериментах алгоритм аналога Topic Tiling фиксирован и отличается только входными данными полученными с первого этапа.

Глава 2

Постановка Задачи

В качестве информации о каждой банковской транзакции имеется идентификатор пользователя, время, дата, сумму транзакции и mcc-код торговой точки (Merchant Category Code) - код категории продавца.

2.1 Векторное представление mcc-кода

А надо ли?? По сути, тематическая модель мне нужна именно для получения векторного представления, но никаких свойств векторного представления особо не использую.

2.2 Тематическая модель: основные понятия и определения

Хотим для любой транзакции получить её представление в виде вектора тем. Для этого построим тематическую модель обученную на транзакционных данных. И достанем из нее информацию о тематике транзакций.

Обозначим D - множество профилей пользователей (коллекция). Пусть W - множество mcc-кодов (словарь). Будем называть профилем пользователя $d \in D$ совокупность дополнительной информации о пользователе, такой как пол, возраст, образование, семейное положение и историю (последовательность) его транзакций $w_{d,1}, \ldots, w_{d,N_d}$, где N_d - число транзакций пользователя за выбранный период.

Предполагаем, что для определения характера потребления пользователя порядок транзакций в его истории не важен, а важен набор mcc-кодов и суммарная трата на каждый код (гипотеза о мешке слов). Обозначим n_{wd} - сумма транзакций клиента d по mcc-коду w. Тогда, можем заключить информацию об истории всех пользователей в матрицу частот mcc-кодов в истории пользователей F размера $|W| \times |D|$ состоящую из элементов n_{wd}.

Можем записать следующие частотные оценки вероятностей, которые мы можем

явно получить зная коллекцию транзакций:

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \quad \hat{p}(d) = \frac{n_d}{n}, \quad \hat{p}(w) = \frac{n_w}{n}, \quad \hat{p}(w|d) = \frac{n_{dw}}{n_d}; \tag{1}$$

$$n_d = \sum_w n_{dw} - \text{общая сумма трат пользователя } d;$$

$$n_w = \sum_d n_{dw} - \text{сумма покупок всех пользователей по mcc-коду} w;$$

$$n = \sum_d \sum_w n_{dw} - \text{сумма трат всех пользователей.}$$

Полагаем, что наличие любого mcc-кода $w \in W$ в истории транзакций пользователя $d \in D$ связано с некой скрытой переменной $t \in T$. Где T - множество возможных составляющих характера поведения пользователя (темы).Тогда можем представить коллекцию D как выборку n_{tdw} из $p(d,w,t)$ на множестве $D \times W \times T$. Где n_{tdw} - связанная с темой t сумма транзакций клиента d по mcc-коду w.

В качестве гипотезы условной независимости возьмем предположение о том, что вероятность наличия mcc-кода в истории пользователя d с характером потребления t зависит от характера t, но не зависит от пользователя d, то есть может быть описана общим для всех пользователей распределением $p(w|t)$:

$$p(w|d,t) = p(w|t). \tag{2}$$

Тогда с помощью формулы полной вероятности и гипотезы условной независимости можем описать распределение mcc-кодов у пользователя $p(w|d)$ вероятностной смесью распределений mcc-кодов в характерах потребления $\phi_{wt} = p(w|t)$ с весами $\theta_{td} = p(t|d)$:

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}. \tag{3}$$

Вероятностная модель (3) описывает процесс порождения истории транзакций по известным распределениям $p(w|t)$ и $p(t|d)$.

Задача тематического моделирования — это обратная задача: по заданному множеству профилей D требуется найти параметры ϕ_{wt} и θ_{td}, при которых тематическая модель (3) хорошо приближает частотные оценки условных вероятностей $\hat{p}(w|d) = \frac{n_{wd}}{n_d}$.

Выпишем частотные оценки вероятностей связанных с характером потребления t:

$$\hat{p}(t) = \frac{n_t}{n}, \quad \hat{\phi}_{wt} = \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \quad \hat{\theta}_{td} = \hat{p}(t|d) = \frac{n_{td}}{n_d}, \quad \hat{p}(t|d,w) = \frac{n_{tdw}}{n_d w}; \tag{4}$$

$$n_{td} = \sum_w n_{tdw} - \text{связанная с темой } t \text{ сумма транзакций клиента } d;$$

$n_{wt} = \sum_d n_{tdw}$ – связанная с темой t сумма транзакций по mcc-коду w;

$n_t = \sum_d \sum_w n_{tdw}$ – сумма трат всех пользователей связанная с темой t.

Зная $\hat{\phi}_{wt}$ и $\hat{\theta}_{td}$ с помощью (1) и (4) можем оценить распределение $p(t|w)$:

$$\hat{p}(t|w) = \frac{p(w|t)p(t)}{p(w)} = \frac{\hat{\phi}_{wt} \sum_d \hat{\theta}_{td} n_d}{n_w}$$

Распределение $p(t|w)$ называют тематикой транзакции w. Для векторного представления транзакции будем использовать её тематику. То есть вектор размерности $|T|$ со значениями $p(t|w)$.

Простейшей вероятностной тематической моделью является модель PLSA [1]. В PLSA для построения модели (3) максимизируется логарифм правдоподобия при ограничениях нормировки и неотрицательности:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \;\to\; \max_{\Phi, \Theta}; \tag{5}$$

$$\sum_{w \in W} \phi_{wt} = 1; \quad \phi_{wt} \geq 0; \qquad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geq 0. \tag{6}$$

Для решения этой задачи воспользуемся EM-алгоритмом — итерационный процесс, состоящий из двух шагов: E-шаг (expectation) и M-шаг (maximization). На E-шаге по текущим параметрам ϕ_{wt} и θ_{td} (начальное приближение — нормированные неотрицательные случайные вектора) вычисляются вероятности $p(t|d, w)$ для всех $t \in T, w \in W, d \in D$. На M-шаге при фиксированных вероятностях $p(t|d, w)$ вычисляются новые приближения для параметров ϕ_{wt} и θ_{td}.

Равенство (3) перепишем в матричном виде. В левой части равенства находится известная матрица частот MCC-кодов у клиентов $F = \big(\hat{p}(w|d)\big)_{W \times D}$. Правая часть представляет собой произведение двух неизвестных матриц — матрицы $\Phi = (\phi_{wt})_{W \times T}$ и матрицы $\Theta = (\theta_{td})_{T \times D}$. Считаем, что $|T|$ много меньше $|D|$ и $|W|$, поэтому задача тематического моделирования сводится к поиску приближённого матричного разложения $F \approx \Phi\Theta$, ранг которого не превышает $|T|$.

2.2.1 Аддитивная регуляризация тематических моделей

Задача стохастического матричного разложения является некорректно поставленной, так как множество её решений в общем случае бесконечно. Если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ также является решением для всех невырожденных матриц S, при условии, что матрицы ΦS и $S^{-1}\Theta$ — стохастические.

Аддитивная регуляризация тематических моделей (ARTM) [2] основана на максимизации линейной комбинации логарифма правдоподобия и нескольких регуляри-

заторов $R_i(\Phi, \Theta),\ i = 1, \ldots, k$:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \ \rightarrow \ \max_{\Phi, \Theta}; \qquad R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta); \qquad (7)$$

при ограничениях (6), где τ_i — неотрицательные коэффициенты регуляризации. Эта задача также решается с помощью EM-алгоритма, изменения по сравнению с PLSA претерпевают формулы M-шага.

2.3 Сегментация транзакционных данных

Сформулируем задачу о сегментации профиля пользователя. Дан профиль пользователя $d \in D$, определенный последовательностью транзакций $w_{d,1}, \ldots, w_{d,n_d}$, даны векторные представления этих транзакций $\bar{t}_w \in \mathbf{R}^m$. Определить границы сегментов профиля $s_{d,1}, \ldots, s_{d,m-1}$, где m - число сегментов в документе d. Так, чтобы они разделяли различные однородные временные участки.

Где в качестве временных участков будем иметь ввиду полуинтервалы во времени или последовательности подряд идущих транзакций внутри выбранного периода.

Для проведения обычной сегментации профиля пользователя (без привлечения тематик транзакций, а используя исходные mcc-коды) будем задавать векторные представления транзакций с помощью one-hot-encoding векторов с единицей на месте идентификатора mcc-кода.

Для проведения тематической сегментацией профиля пользователя будем задавать векторные представления транзакций, как их тематики $p(t|w)$ в тематической модели.

2.4 Оценка качества сегментации

Для оценки качества проведения границ сегментов используются метрики качества сегментации такие как P_k - мера и Window Diff.

2.4.1 P_k - мера

Для каждого профиля рассматриваются пары транзакций:

$$(w_{d,1}, w_{d,k+1}), \ldots, (w_{d,n_d-k}, w_{d,n_d})$$

Для каждой пары записывается 0, если они находятся в одном сегменте и 1, если в разных. P_k - доля несовпадений между значениями оцениваемой и образцовой сегментацией.

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

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

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

2.4.2 Window Diff

В отличие от P_k - меры Window Diff ставит каждой паре транзакций не 1 или 0, а количество границ сегментов, находящихся между транзакциями данной пары. Window Diff - доля несовпадений между значениями оцениваемой и образцовой сегментацией.

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

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

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

Глава 3

Построение моделей, вычислительные эксперименты

3.1 Описание и обработка транзакционных данных

В работе использованы данные о всех транзакциях для набора пользователей за примерно три года. Транзакции записаны в таблице, имеющей поля:

- mcc_code - mcc-код транзакции
- $amount\_ru$ - сумма транзакции в рублях
- $cardnumber$ - уникальный идентификатор карты пользователя
- $trans\_time$ - дата и время транзакции (с точностью до секунды)

Где сумма транзакции указана со знаком, означающим списание или зачисление средств. В эксперименте использовались только данные о списаниях. То есть были взяты только транзакции с $amount\_ru < 0$. Были удалены все транзакции с mcc-кодами финансовых операций и некоторые частые mcc-коды, такие как коды покупок в супермаркетах.

Для некоторых пользователей доступна дополнительная информация об образовании, поле, возрасте, семейном положении.

Для каждого пользователя была выделена и упорядочена во времени последовательность его транзакций (профиль).

3.2 Аналог Topic Tiling

Предлагается решать задачу о проведении сегментов в общем виде с помощью аналога алгоритма Topic Tiling [3].

- На вход подаем последовательность транзакций(профиль), представленных в векторном виде (рис.(1) (a)).

- 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 (рис.(1) (б) cosine_similarity).
- Строим график cos-близости от i и сглаживаем его окном h_2 (рис.(1) (б) smoothed_cosine_similarity).
- отмечаем все локальные минимумы на графике cos-близости - это потенциальные места для проведения границ сегмента.
- Для всех локальных минимумов 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 (рис.(1) (б) depth_score).

- Если количество сегментов задано как n, то выберем $n-1$ максимальных по значению $depth\_score(i)$ мест и проведем в них границы сегментов (рис.(1) (б) segment border).
- Если количество сегментов не задано, то объявляем $threshold = \mu + \frac{\sigma}{2}$ и проводим границы везде, где $depth\_score(i) > threshold$ (рис.(1) (б) segment border).

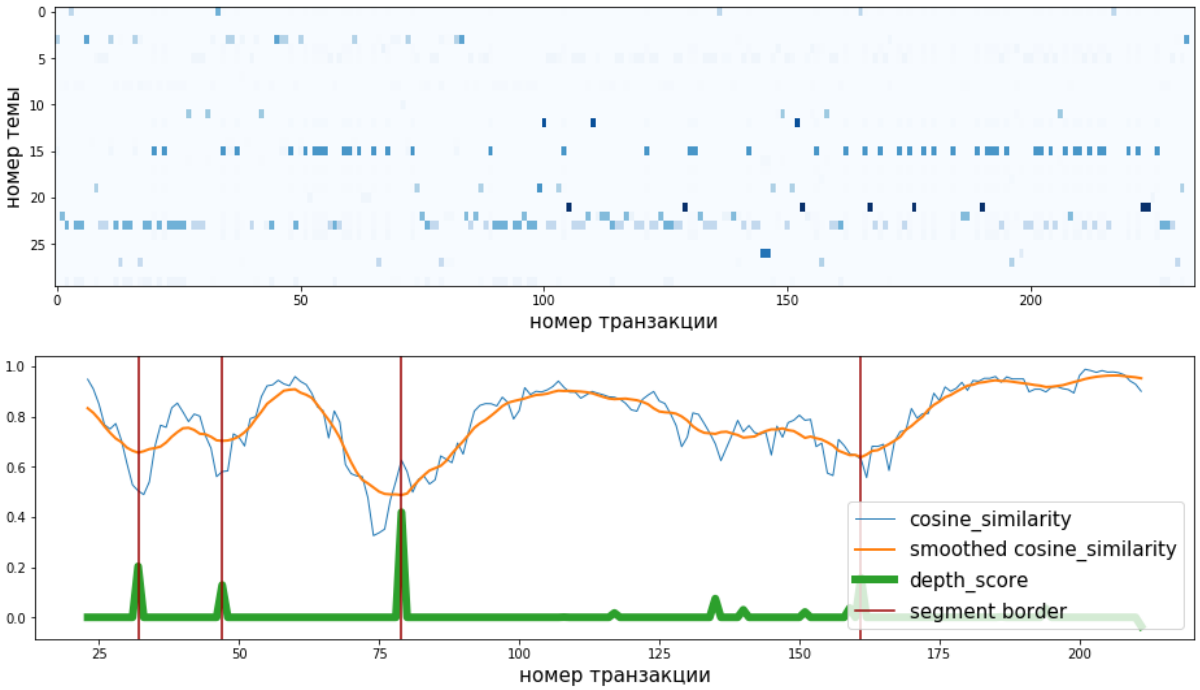

Рис. 1. а) Векторное представление одного профиля (сверху).
б) Стадии работы Аналога Topic Tiling (снизу)

Отличие данного алгоритма от [3] в двух моментах: Первое изменение - добавлено сглаживание графика cos-близости перед нахождением локальных экстремумов.

Второе изменение заключается в изменении порога принятия решения о выставлении границы сегмента.

3.3 Краткие описания моделей

Модели ниже задают векторное представление транзакций с помощью их тематик.

- **PLSA** - PLSA модель без модальностей на 30 темах
- **LDA** - LDA модель без модальностей на 30 темах
- **one-hot** - тематика транзакции заданна one-hot-encoding вектором с единицей на месте идентификатора mcc-кода.
- **random** - тематика транзакции каждого mcc-кода задана вектором, сгенерированным из равномерного распределения.
- **lazy** - тематика всех транзакций одинакова (то есть модель не проводит границы сегментов).
- **PLSA_30_modality** - PLSA модель на 30 темах с модальностями, построенная по подкатегориям товаров.
- **PLSA_50_modality** - PLSA модель на 50 темах с модальностями, построенная по подкатегориям товаров.
- **ARTM_30** - ARTM модель на 30 темах с модальностями (субьективно лучшая по качеству тем)

3.4 Эксперименты

3.4.1 Поиск границ сегментов

Цель данного эксперимента определить качество сегментации искусственных профилей алгоритмом Аналог Topic Tiling для векторных представлений транзакций из различных моделей.

Полагаем, что разница между профилями разных пользователей более значительны, чем разница между частями внутри профиля одного пользователя.

Создадим несколько искусственных профилей, склеив части реально существующих профилей.

Для этого зафиксируем длину искусственного профиля(количеств транзакций). Случайным образом определим места разреза (склеивания). Вставим в эти места части подходящей длины вырезанные из случайных реальных профилей.

Установим истинные границы сегментов на местах склеивания частей различных профилей.

	P_k		WindowDiff	
Модель	mean	std	mean	std
lazy	0.434	0.103	0.434	0.103
one-hot	0.688	0.179	0.570	0.201
random	0.673	0.183	0.543	0.211
LDA	0.700	0.176	0.584	0.196
PLSA	0.705	0.179	0.588	0.199
PLSA_30_modality	0.587	0.181	0.431	0.164
PLSA_50_modality	0.601	0.172	0.452	0.168
ARTM_30	0.676	0.173	0.549	0.198

Таблица 1. Точность проведения сегментов по сравнению с истинными на 300 искусственных профилях

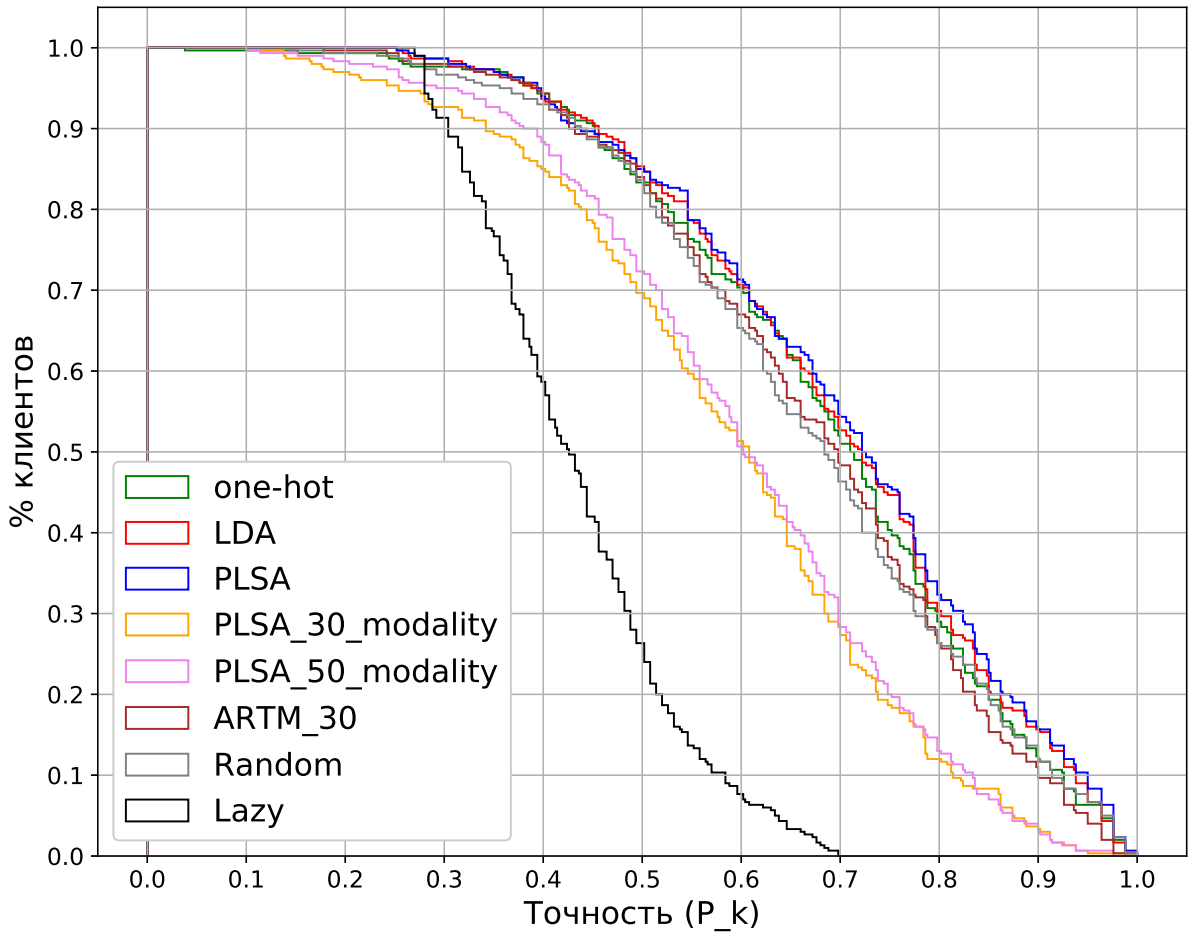

Рис. 2. Точность проведения сегментов по сравнению с истинными на 300 искусственных профилях

В таблице (1) и на графике (2) представлена точность проведения сегментов по сравнению с истинными границами сегментов на искусственной выборке для различных моделей генерации векторных представлений. Точность измерена с помощью

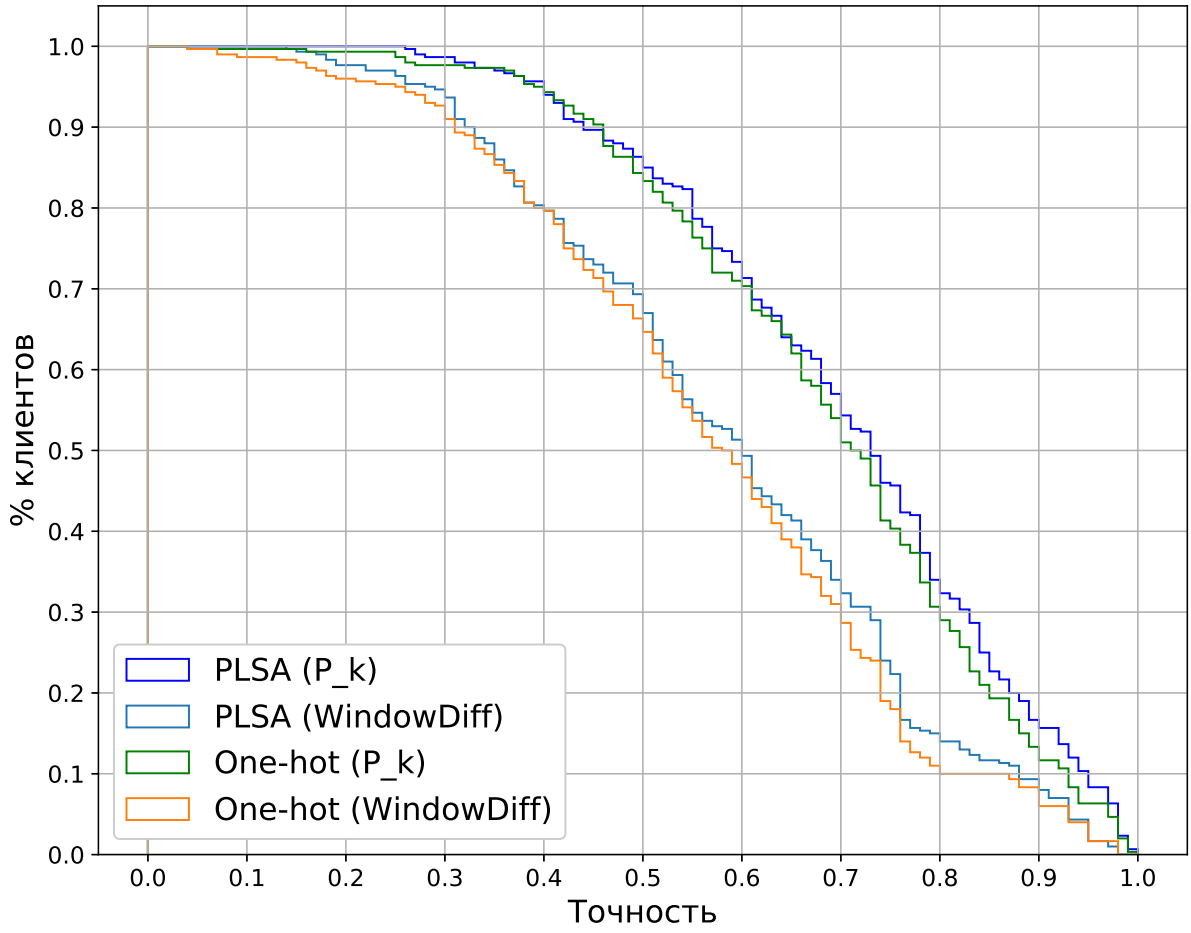

Рис. 3. Сравнение P_k и WindowDiff меры измерения качества

P_k и WindowDiff меры для 300 искусственных профилей. В таблицу занесены mean (выборочное среднее) и std (среднеквадратичное отклонение) точности.

На график (2) изображены кривые процентного отношения пользователей, для чьих профилей точность (P_k) сегментации не хуже заданной. Из таблицы (1) и графика (2) видно, что лучшее качество(P_k) у модели PLSA (наибольший mean, также на большей части оси график PLSA самый верхний), не значительно меньше качество у LDA и one-hot. На графике (3) показано сравнение P_k и WindowDiff метрик измерения качества для моделей PLSA и One-hot. Из него и таблицы (1) видно, что они одинаково ранжируют модели по качеству проведения сегментов. **Сомнения** Из таблице (1) и графика (2) видно, что более сложные модели (ARTM_30, PLSA_50_modality, PLSA_30_modality) уступают по качеству более простым (PLSA, LDA).

Сомнения(А надо ли? вопрос в полезности и дизайне эксперимента) факт про одинаковость one-hot и random представления кодов в смысле распределений точности определения синтетических сегментов. + стат тесты на гипотезу однородности распределения ошибок.

В таблице (2) представлена точность (P_k) в зависимости от длины синтетического профиля (количество транзакций). Видно, что с увеличением длины профиля

точность (P_k) точность сегментации возрастает.

Сомнения(WindowDiff тоже возрастает отобразить ли это?. Также в обоих случаях возможны искажающие зависимость от длинны эффекты из-за методов выбора окон в Topic Tiling и оценки качества, они хоть и динамичные от длины документа, но различны.

	One-hot(P_k)		LDA(P_k)	
Длина профиля	mean	std	mean	std
100	0.696	0.174	0.722	0.174
200	0.764	0.157	0.759	0.159
500	0.805	0.147	0.802	0.150
1000	0.806	0.150	0.808	0.150

Таблица 2. Точность проведения сегментов в зависимости от длины документа

3.4.2 Сравнение простого и тематического сегментирования

Целью этого эксперимента является определить схожесть простой сегментации (модель One-hot) и тематической сегментации.

Сравним границы сегментов проведенные простым и тематическим сегментированием на реальных профилях. Найдем WindowDiff и P_k меры, определяя простую сегментацию как истинно верную. Занесем результаты в таблицу (3) и график (4).

Видим, что наиболее похоже на One-hot по P_k сегментирование с помощью модели LDA, незначительно меньше похоже сегментирование с помощью PLSA.

	P_k		WindowDiff	
Модель сегментации	mean	std	mean	std
lazy	0.517	0.308	0.517	0.308
random	0.820	0.176	0.707	0.238
LDA	0.848	0.174	0.758	0.236
PLSA	0.845	0.172	0.743	0.233
PLSA_30_modality	0.646	0.240	0.535	0.275
PLSA_50_modality	0.668	0.230	0.551	0.272
ARTM_30	0.744	0.207	0.625	0.261

Таблица 3. Результаты сравнения тематического сегментирования с простым на 1000 реальных профилях

Скачек точности на единице на графике (4) объясняется тем, что примерно у 20% клиентов One-hot модель не провела ни одной границы сегмента.

3.5 Выводы

Сомнения

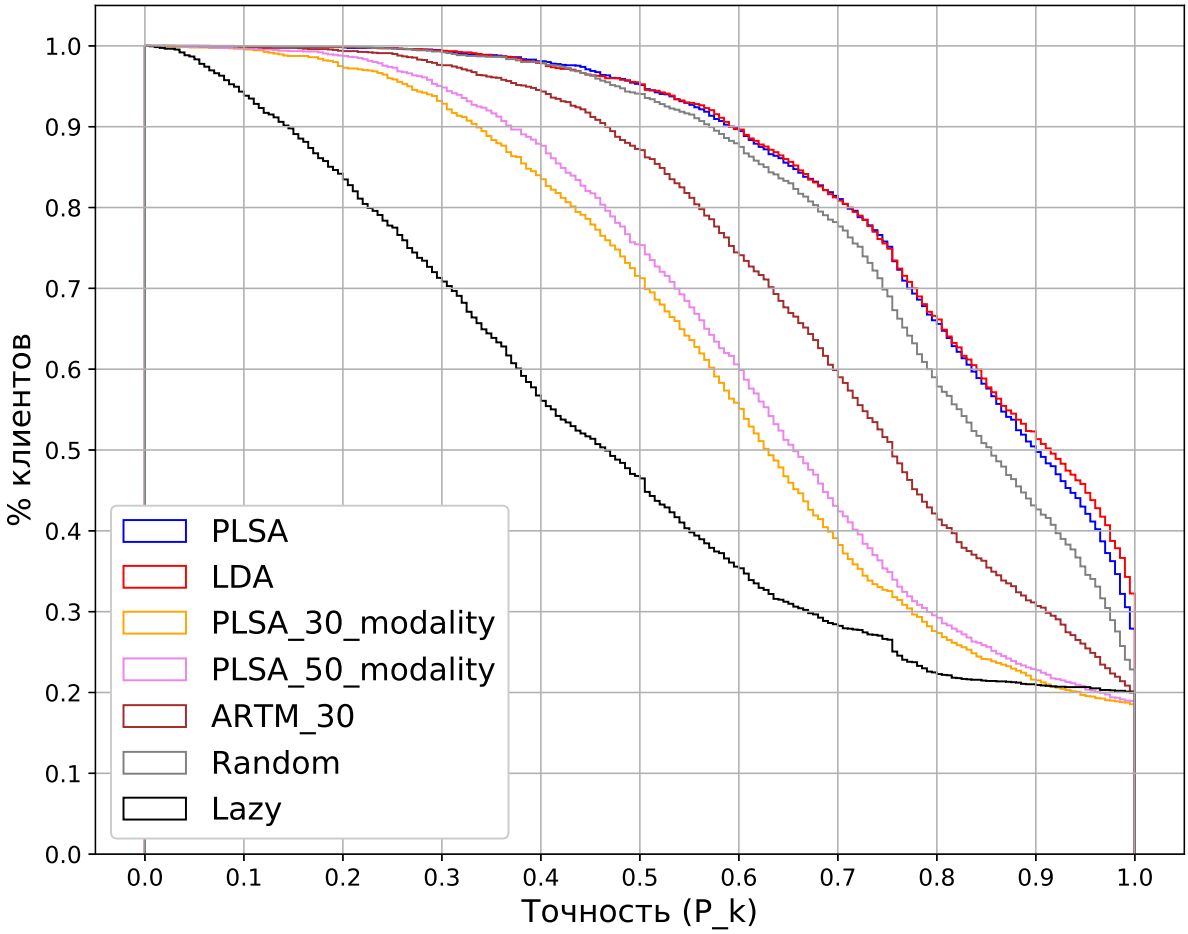

Рис. 4. Результаты сравнения тематического сегментирования с простым на 1000 реальных профилях

1) простые модели справляются лучше всего

2) случайное векторное представление справляется тоже неплохо

3) хорошие модели сегментируют хуже

4) более хорошие (ARTM_30) немного лучше

5) точность(на искусственных профилях) увеличивается с увеличением длины профиля

6) для ранжирования моделей по P_k и WindowDiff мере годится любая из них.

Глава 4

Заключение

Литература

[1] Thomas Hofmann. Hofmann, t.: Unsupervised learning by probabilistic latent semantic analysis. machine learning 42(1-2), 177-196. 42:177–196, 01 2001.

[2] K. V. Vorontsov. Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3):301–304, May 2014.

[3] Martin Riedl and Chris Biemann. Text Segmentation with Topic Models . *Journal for Language Technology and Computational Linguistics (JLCL)*, 27(47-69):13–24, 2012.