Сегментирование профиля пользователя по транзакционным данным

Козлинский Е.М.

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

24.04.19

План

Цель работы

Мотивация

- Выяснить, влияет ли представление профиля пользователя с помощью тематического моделирования на возможность выделить в нем сегментную структуру.
- Хотим получить внешний критерий качества модели тематического моделирования, основанный на сохранении сегментирующей способности.

Задача

Выделение сегментной структуры во временном ряде истории пользователя, полученное с помощью тематического моделирования.

Входные данные

Данные

11 миллионов транзакций, содержащих
идентификатор пользователя, дату, сумму транзакции, mcc-код

Предобработка

- Удаление кодов, связанных с финансовыми операциями
- сортировка по времени
- группировка по клиентам (составление профиля пользователя)

Постановка задачи

- Определим профиля пользователя $d \in D$ последовательностью транзакций $w_{d,1}, \ldots, w_{d,n_d}$, где n_d - число транзакций пользователя за выбранный период. Определим множество тем T коллекции D.
- Задача тематической сегментации - определить для каждой транзакции $w \in d$, для каждого профиля $d \in D$ вектор тем $\bar{t} \in \bar{T}$. И найти для каждого профиля d границы между монотематическими сегментами $s_{1,d}, \ldots, s_{m,d}$, где m - число сегментов в профиле d.

Аналог Topic Tiling

- Проходим по профилю двумя скользящими окнами $(i-h, h]$ и $(h, i+h]$, для каждого i вычисляя cos-близость между средним левого и правого окна.
- Строим график cos-близости от i и отмечаем все локальные минимумы - 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.

Аналог Topic Tiling

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

Алгоритм №2

- сглаживаем профиль с помощью скользящего окна h_1
- находим срезы профиля с периодом h_2
- строим график косинусных расстояний между срезами
- находим $depth\_score$ по аналогии с Topic Tiling
- варьируя h_1, h_2 и суммируя $depth\_score$, получаем график уверенности проведения сегментов
- для каждого "пика"на графике определяем наилучшую точку для проведения границы сегмента

Оценка качества

Промежуточная оценка качества

Смотрим на корреляцию Кендела и Фехнера между двумя графиками степени уверенности проведения сегментов.

Итоговая оценка

Качество сегментации профиля оцениваем сравнивая найденные границы сегментов с границами сегментации, найденной по сырым данным с помощью метрик качества сегментации P_k и Window Diff.

P_k-мера

Для каждого профиля рассматриваются пары транзакций:

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

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.

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

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

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

Topic Tiling

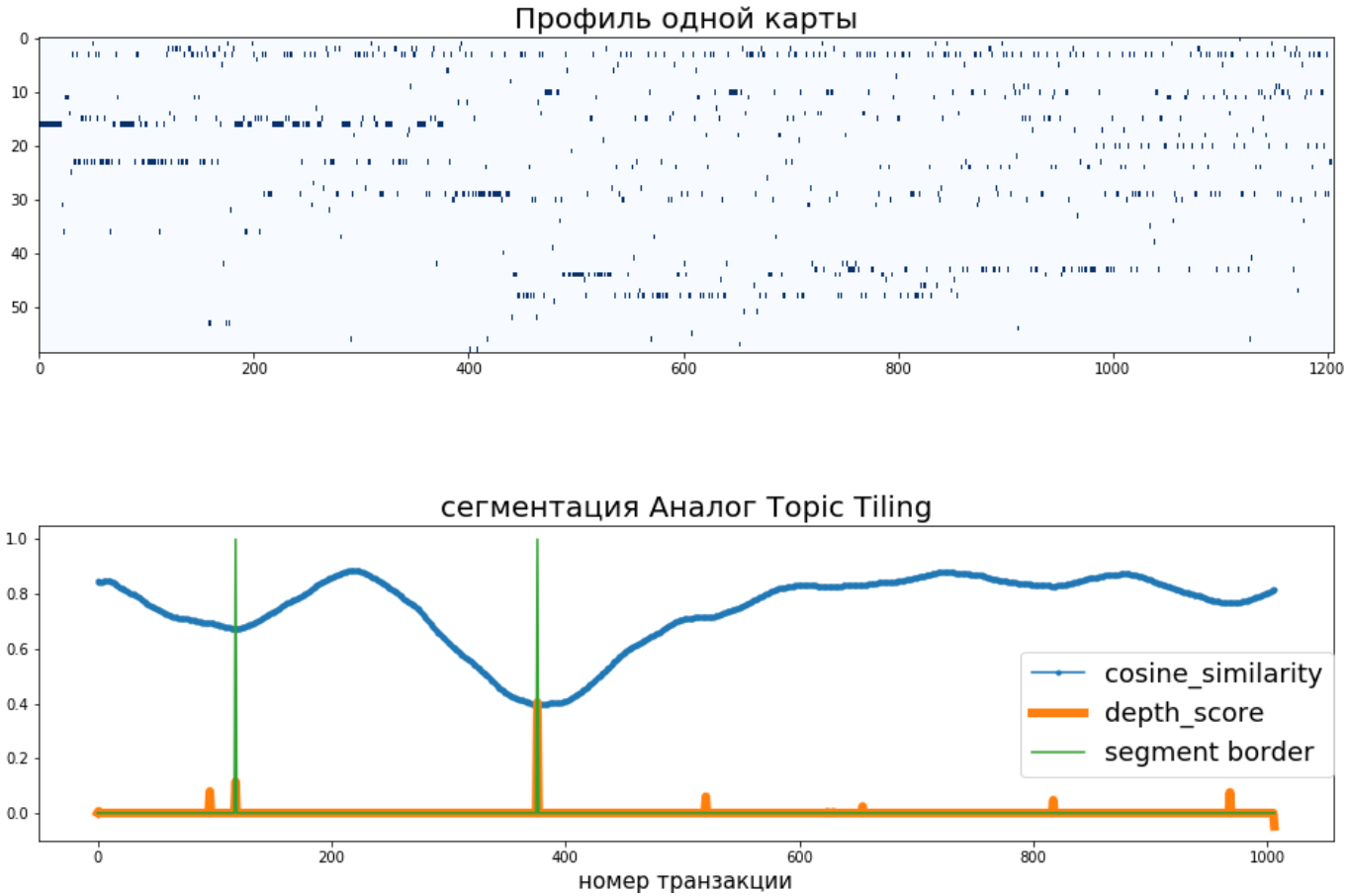

Рис.: Сегментирование профиля по последовательности mcc-кодов

Алгоритм №2

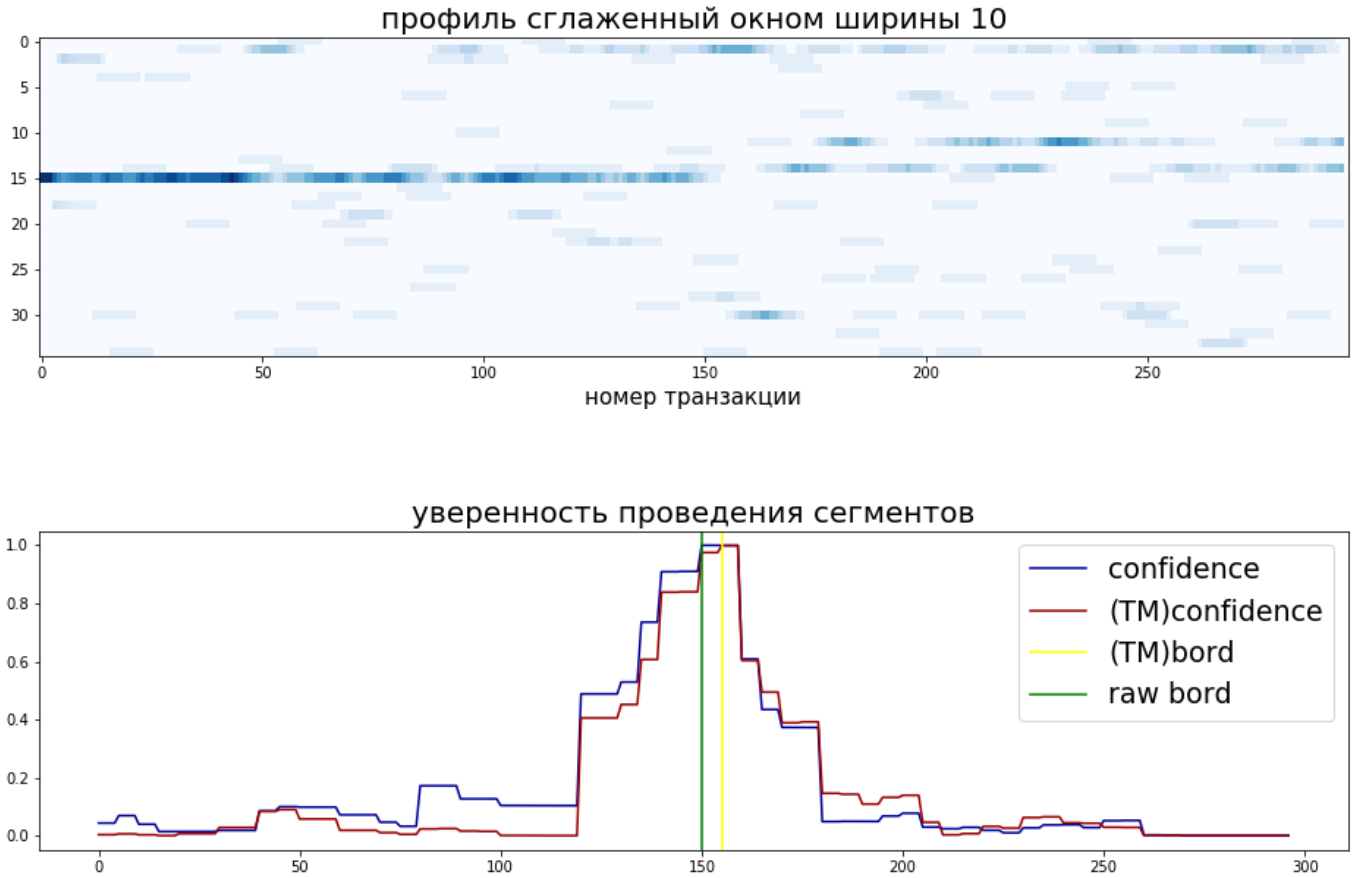

Рис.: Уверенность сегментации и проведение сегментов

Заключение

На данный момент:

- Разработан алгоритм №2 для построения сегментации
- Статистически показано присутствие корреляции между графиками уверенности сегментации для алгоритма №2
- Разработан алгоритм Аналог Topic Tiling для построения сегментации

Планы

- Реализовать сравнение построенных сегментаций используя P_k меру
- Улучшить качество алгоритма Аналог Topic Tiling
- Построить внешний критерий качества тематической модели на основе качества сегментирования с ее помощью

Литература