Министерство науки и высшего образования Российской Федерации
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(государственный университет)

Факультет управления и прикладной математики

Базовая организация — ФИЦ ИУ РАН,
Вычислительный центр им. А.А. Дородницына РАН

Кафедра «Интеллектуальные системы»
специализация «математические и информационные технологии (ПиОС)»

Квалификационная работа на соискание степени бакалавра

по направлению 03.03.01 «Прикладные математика и физика»,

профиль «Компьютерные технологии и интеллектуальный анализ данных»

Применение мультимодальных тематических моделей

к анализу транзакционных данных банков

Студент группы 473б — Никитин Филипп Александрович

Научный консультант
д.ф-м.н. — Воронцов Константин Вячеславович

Научный руководитель
д.т.н. — Матвеев Иван Алексеевич

Москва, 2018

Содержание

Аннотация

Рассматривается проблема построения профиля потребления клиента по банковским транзакционным данным. Для решения проблемы используется тематическое моделирование: строились модели PLSA , LDA, ARTM, исследовались зависимости результата от различных регуляризаций. Полученные результаты свидетельствуют о применимости тематического моделирования к поставленной задаче. Разработанные методы позволяют строить интерпретируемые профили потребления по данным о транзакционной активности и дополнительным метаданным клиентов банка и улучшать их интерпретируемость без ухудшения правдоподобия.

1 Введение

Работа посвящена созданию модели построения скрытых профилей потребления клиентов по транзакционным данным банковских карт с использованием методов мягкой кластеризации, основанных на LDA [1], PLSA [2], подходе ARTM [3].

Ставится задача по созданию способа построения скрытых профилей потребления клиентов по транзакционным данным банковских карт. Модель ставит в соответствие каждому клиенту набор типов потребления, одновременно вычисляя степень принадлежности клиента каждому из них. Тип потребления описан входящими в его состав категориями покупок и распределениями денежного потока по этим категориям. Важным условием является интерпретируемость построенных типов потребления.

Построение профилей потребления клиентов способствует выделению групп пользователей с одинаковым шаблоном поведения. Также выделение профилей потребления и анализ степени покрытия ими выборки позволят выделить группы пользователей, целевая работа с которыми наиболее эффективна. Агрегированная статистика по близким клиенту людям позволит пользователю оценить потенциал собственных возможностей и уровень успеха своей деятельности, а банку — построить стратегию повышения этого потенциала для конкретного пользователя.

Рассматриваемая задача возникает во многих прикладных областях, например, при построении рекомендательных систем для web ресурсов [4, 5].

Наряду с моделями ARTM, LDA, ARTM для решения этого типа задач используются классические методы кластеризации — метод k-средних, иерархическая кластеризация и другие [6].

В области анализа текстов существует аналогичная задача выявления тематики коллекции документов [3]. Данную задачу решают с помощью вероятностного тематического моделирования (probabilistic topic modeling), в результате которого каждый документ описывается дискретным распределением вероятностей тем, а каждая тема — дискретным распределением вероятностей слов. В основе вероятностного тематического моделирования лежит гипотеза условной независимости, а также гипотеза "мешка слов"состоящая в том, что для определения тематики документа порядок слов не имеет значения. Данная гипотеза обеспечивает интерпретируемость и разреженность компонент в тематических векторных представлениях слов.

Таким образом, поставленная задача выделения типов потребления может быть решена с помощью тематического моделирования, если понятию «документ» сопоставлен клиент, понятию «слово» — категория покупки, а понятию «тема» — тип потребления.

Целью работы является выявление типов потребления клиентов с помощью тематической модели, а также сравнение нескольких подходов ее построения с точки зрения интерпретируемости полученных типов потребления.

2 Постановка задачи

2.1 Задача тематического моделирования на транзакционных данных

Атрибутом каждой банковской транзакции является тройка: идентификатор клиента, совершившего транзакцию, сумма транзакции и MCC-код. MCC-код (Merchant Category Code) — код вида торговой точки, который представляет собой четырехзначный номер и применяется в отрасли банковских карт для классификации торгово-сервисных предприятий по типу их деятельности.

Пусть D — множество клиентов, W — множество (словарь) категорий покупок. Введем N — матрицу размера $|W| \times |D|$, в которой элемент n_{wd} обозначает сумму транзакций с MCC-кодом w у клиента d. Каждый клиент $d \in D$ описывается последовательностью сумм расходов по MCC-кодам n_{wd}.

Выборка представляет собой коллекцию записей о транзакциях клиента.

Предполагается, что появление каждого MCC-кода w у клиента d связано с некоторой скрытой переменной t из конечного множества типов потребления T . Тогда коллекция D представляет собой выборку троек (d, w, t), взятых независимо из дискретного распределения $p(d, w, t)$ на множестве $D \times W \times T$.

В качестве гипотезы условной независимости возьмем предположение, что появление MCC-кодов в типе потребления t не зависит от клиента:

$$p(w \mid t) = p(w \mid d, t). \tag{2.1}$$

Согласно формуле полной вероятности и гипотезе условной независимости, распределение MCC-кодов у клиента $p(w \mid d)$ описывается вероятностной смесью распределений MCC-кодов в типах потребления $\varphi_{wt} = p(w \mid t)$ с весами $\theta_{td} = p(t \mid d)$:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t, d)\, p(t \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d) = \sum_{t \in T} \varphi_{wt}\theta_{td}. \tag{2.2}$$

Вероятностная модель (2.2) описывает процесс порождения транзакций по известным распределениям $p(w \mid t)$ и $p(t \mid d)$. Задача тематического моделирования — это обратная задача: по заданной коллекции D требуется найти распределения φ_{wt} и θ_{td}.

Равенство (2.2) перепишем в матричном виде. В левой части равенства находится известная матрица частот MCC-кодов у клиентов $N = \big(\hat{p}(w \mid d)\big)_{W \times D}$.

Правая часть представляет собой произведение двух неизвестных матриц — матрицы $\Phi = (\varphi_{wt})_{W \times T}$ 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 $N \approx \Phi\Theta$, ранг которого не превышает $|T|$.

Матрица Φ описывает тип потребления t как взвешенный набор φ_t MCC-кодов, по которому можно судить об интерпретируемости типа потребления и о денежном потоке по этим MCC-кодам. Матрица Θ позволяет описать профиль потребления θ_d пользователя d как взвешенную смесь типов потребления.

2.2 Вероятностный латентно-семантический анализ

Первая тематическая модель была предложена Томасом Хофманном в 1999 году [2]. Для решения поставленной задачи используется метод максимума правдоподобия. Данный метод применяется в математической статистике для оценки неизвестных параметров вероятностных моделей по полученным данным. Функция правдоподобия определяет вероятность полученной выборки от параметров. В предположении о независимости наблюдений она может быть представлена в следующем виде:

$$\prod_{i=1}^{n} p\,(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(w \mid d)^{n_{wd}} p(d)^{n_{wd}} \;\to\; \max_{\Phi,\Theta} \tag{2.3}$$

Пролагарифмировав правдоподобие перейдём от произведения к сумме:

$$\sum_{d \in D} \sum_{w \in d} n_{wd} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + n_{wd} \ln(p(d)) \to \max_{\Phi,\Theta}. \tag{2.4}$$

$$\sum_{w \in W} \varphi_{wt} = 1; \quad \varphi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0. \tag{2.5}$$

Для нахождения точки локального экстремума задачи (2.4) воспользуемся условиями Каруша-Куна-Таккера, полученная из них система уравнений,

задающая стационарную точку имеет вид:

$$\begin{cases} p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \\ \varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum\limits_{d\in D} n_{dw}p_{tdw}\Big); \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum\limits_{w\in d} n_{dw}p_{tdw}\Big). \end{cases} \tag{2.6}$$

Здесь введен оператор norm, который преобразует произвольный заданный вектор $(x_i)_{i\in I}$ в вектор вероятностей $(p_i)_{i\in I}$ дискретного распределения путем обнуления отрицательных элементов и нормировки:

$$p_i = \underset{i\in I}{\operatorname{norm}}(x_i) = \frac{(x_i)_+}{\sum\limits_{j\in I}(x_j)_+}, \text{ для всех } i \in I, \tag{2.7}$$

где $(x)_+ = \max\{0, x\}$ — операция положительной срезки или неотрицательного нормирования.

Для решения системы (2.6) воспользуемся EM-алгоритмом [7]. Это 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 $p(t \mid d, w)$ для всех $t \in T, w \in W, d \in D$. На M-шаге при фиксированных вероятностях $p(t \mid d, w)$ вычисляются новые приближения для параметров φ_{wt} и θ_{td}.

2.3 Латентное размещение Дирихле

В рассмотренной выше модели PLSA предполагается, что данные порождаются вероятностной моделью с параметрами Φ, Θ, однако можно предположить, что параметры Φ, Θ также являются случайными и подчиняются априорному распределению $p\,(\Phi, \Theta, \gamma)$, где γ — гиперпараметр распределения. В этом случае максимизация совместного правдоподобия данных и модели приводит к задаче максимума апостериорной вероятности:

$$p(D, \Phi, \Theta; \gamma) = p(D \mid \Phi, \Theta)p(\Phi, \Theta; \gamma) = p(\Phi, \Theta; \gamma)\prod_{i=0}^{n} p(d_i, w_i \mid \Phi, \Theta) \;\to\; \max_{\Phi,\Theta,\gamma} \tag{2.8}$$

Для оптимизации данного функционала перейдем от произведения к сумме логарифмированием.

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + \ln p(\Phi,\Theta;\gamma) \to \max_{\Phi,\Theta;\gamma}. \tag{2.9}$$

Заметим, что данный функционал отличается от модели PLSA (2.4) слагаемым $\ln p(\Phi,\Theta;\gamma)$, то есть имеет дополнительное ограничение на параметры модели.

Также применяется принцип максимизации неполного правдоподобия, параметр γ оптимизируется после интегрирования по случайным параметрам (Φ,Θ). Данный приём позволяет снизить размерность задачи и уменьшить переобучение модели. Для прикладных целей необходимо знать матрицы (Φ,Θ), а не их распределение. Получаемые данным методом оценки этих матриц мало отличаются от оценок полученных из оптимизации 2.9.

Дэвид Блэй, Эндрю Ын, Майкл Джордан являются авторами модели LDA [1] более известную как Latent Dirichlet allocation. Она основана на предположении, что столбцы матриц (φ_t,θ_d) являются случайными векторами, порождаемыми распределением Дирихле с параметрами: $\alpha\in\mathbb{R}^{|T|}$, $\beta\in\mathbb{R}^{|W|}$.

Функция плотности распределения Дирихле выглядит следующим образом:

$$Dir(\theta_d;\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t\Gamma(\alpha_t)}\prod_t\theta_{td}^{\alpha_t-1},\ \alpha_t>0,\ \alpha_0=\sum_t\alpha_t,\ \theta_{td}>0,\ \sum_t\theta_{td}=1 \tag{2.10}$$

$$Dir(\varphi_t;\beta) = \frac{\Gamma(\beta_0)}{\prod_w\Gamma(\beta_w)}\prod_w\varphi_{wt}^{\beta_w-1},\ \beta_w>0,\ \beta_0=\sum_t\beta_t,\ \varphi_{wt}>0,\ \sum_w\varphi_{wt}=1$$

В таком случае аддитивная добавка в выражении для логарифма правдоподобия (2.9) запишется в следующем виде:

$$R(\Phi,\Theta) = \ln\prod_{t\in T}Dir(\varphi_t;\beta)\prod_{d\in D}Dir(\theta_d;\alpha) + const = \tag{2.11}$$

$$\sum_{t\in T}\sum_{w\in W}(\beta_w-1)\ln(\varphi_{wt}) + \sum_{d\in D}\sum_{t\in T}(\alpha_t-1)\ln(\theta_{td}) \tag{2.12}$$

2.4 Аддитивная регуляризация тематических моделей

Вернемся к задаче PLSA (2.2). Данная задача поставлена некорректно, так как её решение не единственно. Пусть $\Phi\Theta$ — решение, но тогда решением также является и $\Phi'\Theta'$, где $\Theta' = S^{-1}\Theta$, $\Phi' = \Phi S$ для любой невырожденной матрицы S. Для проблемы с некорректно поставленными задачами применяют метод под названием регуляризация.

Подход аддитивной регуляризации тематических моделей (ARTM, additive regularization of topic models) [8] основан на максимизации линейной комбинации логарифма правдоподобия и ограничений-регуляризаторов, заданных функцией $R(\Phi, \Theta)$. Оптимизируемая функция в данном случае принимает следующий вид:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}. \tag{2.13}$$

Тогда система уравнений для задачи (2.13) имеет вид:

$$\begin{cases} p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \\ \varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Big); \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big). \end{cases} \tag{2.14}$$

Наложение дополнительных ограничений, регуляризаторов, открывает широкий спектр возможностей для качественного улучшения результатов в различных прикладных задачах [9]. В частности видно, что модель ARTM является обобщением тематических моделей PLSA и LDA: модель ARTM без регуляризаторов тождественна модели PLSA, модель LDA реализуется использованием в качестве регуляризатора функции (2.11).

Ниже приведены регуляризаторы, используемые в данной работе:

1. Регуляризатор сглаживания и разреживания матриц Φ, Θ:

$$R(\Phi, \Theta) = \sum_{t\in T}\sum_{w\in W^m} \beta_{wt} \ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T} \alpha_{td} \ln \theta_{td} \to \max. \tag{2.15}$$

   В данном выражении α_{td}, β_{wt} — коэффициенты регуляризации, положительные значения которых соответствуют сглаживанию, отрицательные — разреживанию. С помощью данного регуляризатора можно реализовать требование о разреженности матрицы Θ, что соответствует

гипотезе о том, что у пользователя проявляется малое количество типов поведения.

2. Декоррелирование матрицы Φ:

$$R(\Phi, \Theta) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus \{t\}} cov(\varphi_t, \varphi_s) \to \max. \tag{2.16}$$

где τ — коэффициент регуляризации. Это ограничение реализует требование различности профилей потребления.

Тематические модели с правильно подобранными коэффициентами регуляризации превосходят модель LDA на текстовых коллекциях [10].

2.5 Мультимодальная тематическая модель

Теперь рассмотрим случай, когда имеются дополнительные признаки, описывающие клиента. Например, семейное положение, пол или возраст. В этом случае профили потребления, полученные по тратам клиента на различные MCC-коды, могут быть уточнены благодаря учету этих признаков с различными весовыми коэффициентами. Такие тематические модели называются мультимодальными [9].

Пусть M — множество независимых наборов признаков, модальностей, W^m — множество признаков, соответствующее модальности m. В качестве оптимизируемой функции рассмотрим взвешенную с коэффициентами λ_m линейную комбинацию log-правдоподобий по модальностям:

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \tag{2.17}$$

Соответственно, система уравнений, задающая решение данной задачи оптимизации принимает вид.

$$\begin{cases} p_{tdw} = \underset{t \in T}{\text{norm}}\big(\varphi_{wt} \theta_{td}\big); \\ \varphi_{wt} = \underset{w \in W^m}{\text{norm}}\Big(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\Big), \quad n_{wt} = \sum\limits_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw}; \\ \theta_{td} = \underset{t \in T}{\text{norm}}\Big(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\Big), \quad n_{td} = \sum\limits_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw}. \end{cases} \tag{2.18}$$

2.6 Использование профилей потребления для решения задачи классификации

Полученные с помощью тематической модели профили потребления $p(t \mid d)$ могут быть использованы для прогнозирования признаков, таких как пол, семейное положение. Для категориальных признаков, может быть поставлена задача классификации, которая формализуется следующим образом:

Пусть Θ — множество профилей клиентов, Y — конечное множество признаков клиентов. Существует целевое отображение заданное функцией $y^* : \Theta \to Y$, значения которого известно на конечном количестве объектов обучающей выборки $\Theta^m = \{(\theta_1, y_1), \ldots, (\theta_m, y_m)\}$. Требуется построить функцию: $a : \Theta \to Y$, аппроксимирующую исходную неизвестную зависимость y^* по заданному критерию качества или функции потерь.

В работе поставленная задача классификация решалась с помощью методов логистической регрессии [11] и градиентного бустинга [12].

2.7 Метрики качества

Выделение метрик качества для исследуемой задачи является проблемой.

Основным критерием с точки зрения коммерческих приложений является интерпретируемость полученных профилей потребления, это требование не может быть однозначно сведено к математической задаче. Степень интерпретируемости может быть установлена в результате опроса экспертов.

Классические метрики качества тематических моделей делятся на внешние и внутренние. Внутренние метрики качества — критерии качества тематических моделей на выделенной выборке. Внешние — оценивают качество модели для конкретной прикладной задачи. Самым известным внутренним критерием является перплексия, она определяется для каждой модальности отдельно:

$$perp_m = \exp\left(-\frac{1}{n_m}\sum_{d\in D}\sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td}\right) \tag{2.19}$$

где $n_m = \sum_{d\in D}\sum_{w\in W^m} n_{dw}$.

Также оценить качество полученных профилей можно с помощью прогнозирования по ним дополнительных, известных признаков клиентов, таких как пол, возраст. В данном случае тематические вектора θ_d являются признака-

ми для классификатора. В этом случае могут быть применены стандартные метрики качества для задачи классификации, такие как полнота, точность, F-мера. Рассмотрим задачу бинарной классификации. Пусть TP — количество истинно-положительных решений, TN — количество истинно отрицательных решений, FP — количество ложно-положительных решений, FN — количество ложно-отрицательных решений. Тогда формулы для точности, полноты и F_β-меры принимают следующий вид:

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FP} \quad F_\beta = (\beta^2 + 1)\frac{PR}{\beta^2 P + R} \tag{2.20}$$

где β — действительный параметр в диапазоне $0 < \beta < 1$, если точность важнее, чем полнота, иначе $\beta > 1$.

В случае, когда классов больше двух (мультиклассовая задача классификации), метрики вычисляются внутри каждого класса, а затем усредняются.

3 Разработка моделей, вычислительные эксперименты

3.1 Описание используемых данных

Исследования проводились на реальных данных. Данные представляют собой таблицу включающую следующие ключевые поля:

1. *client_id* — идентификатор клиента;
2. *mcc_code* — MCC-код транзакции;
3. *trans_date* — дата транзакции;
4. *amount_rur* — сумма транзакции в рублях.

Ниже приведены некоторые основные характеристики данных:

1. количество транзакций — 232 миллиона;
2. количество клиентов — 400 тысяч;
3. количество уникальных MCC-кодов — 548;
4. временной интервал — 2014-2017 года.

Пример транзакционных данных приведен в таблице 1.

Таблица 1. Пример транзакционных данных

	client_id	trans_date	amount_rur	iso_cntry	trans_crncy	mcc_code	trans_type
0	188864383	2014-01-30	3000.0	RU	RUR	6532	7070
1	328329950	2014-01-30	332.0	None	RUR	None	7000
2	154267144	2014-01-30	-30.0	None	RUR	None	3200
3	71428269	2014-01-30	-5.0	RU	RUR	4829	4051
4	71428269	2014-01-30	150.0	RUR	RUR	6011	7010

Для части пользователей были известны некоторые метаданные, среди которых основное значение имели пол, возраст. Также была доступна информация о семейном положении и образовании, однако в данных графах преобладают пропуски, что усложняет использование данных признаков. Краткая характеристика покрытия метаданными исходной выборки:

1. Пол известен для 97% клиентов;

2. Значение возраста известно для 97% клиентов;

3. Информация об образовании присутствует у 18% клиентов;

4. Информация о семейном положении присутствует у 20% клиентов.

Пример метаданных пользователей представлен в таблице 2.

Таблица 2. Пример атрибутов клиентов

	client_id_way4	birth_dt	gender	edu_stts_name	martal_stts_desc
0	188864383	1948-06-20	F	среднее специальное	вдова
1	1003223820	1994-09-28	M	None	None
2	105880415	1971-09-24	F	среднее специальное	замужем
3	107620354	1980-11-17	M	высшее	холост

3.2 Построение базовых моделей PLSA, LDA

В качестве базовых моделей строились унимодальная нерегуляризованная модель PLSA и модель LDA. Количество различных профилей потребления определялось экспериментально: рассматривались модели с количеством типов потребления от 25 до 50. Наилучший результат определялся исходя из интерпретируемости полученных типов потребления. Все дальнейшие модели были построены для 30 типов потребления.

Аналогично работе с текстами из рассматриваемой выборки были отброшены:

1. Клиенты у которых на 3 MCC-кода приходится 97% и более всех транзакций, данные клиенты имеют вырожденный профиль потребления, для выделения которого не имеет никакого смысла применять сложные алгоритмы.

2. MCC-коды, встречающиеся с высокой частотой у большинства клиентов. В данной категории было отсечено 2 типа MCC-кодов: снятие наличных, перевод физическим и юридическим лицам. Данная операция необходима, так как модели максимизирующей правдоподобие будет выгодно в каждом профиле потребления использовать данные MCC-коды.

Графики зависимости перплексии от номера итерации представлены на рис 1. Качество получившихся профилей неоднозначное: просматриваются как интерпретируемые типы потребления (таблица 3), так и сочетания MCC-кодов, не интерпретируемых однозначно.

Рис. 1. Графики перплексии для модели PLSA и LDA

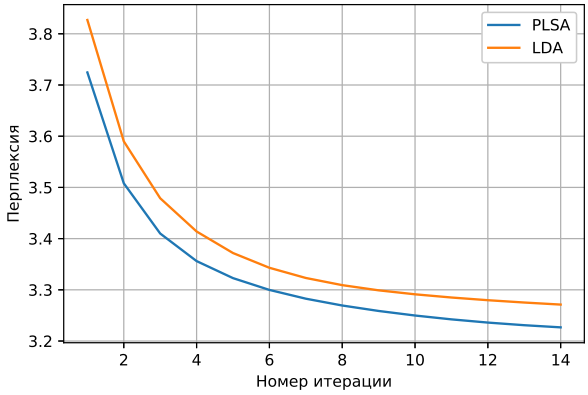


Таблица 3. Интерпретируемые профили потребления для модели PLSA и LDA

φ_{wt}	**MCC-код**	**Описание MCC-кода**
	PLSA	
0.60	3000	Авиалинии, авиакомпании
0.09	4722	Туристические агентства и организаторы экскурсий
0.02	4112	Пассажирские железнодорожные перевозки
	LDA	
0.39	5541	Станции техобслуживания
0.19	5533	Автозапчасти и аксессуары
0.08	5983	Горючее топливо - уголь, нефть, разжиженный бензин, дрова
0.05	6300	Продажа страхования, гарантированное размещение, премии
0.05	5411	Бакалейные магазины, супермаркеты
0.04	7538	СТО общего назначения
0.02	5532	Автошины

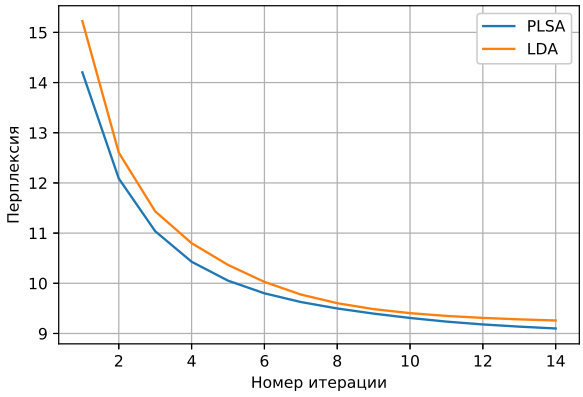
Результаты работы базовой модели показывают, что тематические модели

могут выделять в транзакционных данных интерпретируемые профили. Модель структурирует близкие с точки зрения поведения MCC-коды. С другой стороны существуют профили потребления, которые практически дублируют друг друга, но выражены через разные MCC-коды.

Для решения этой проблемы была разработана группировка MCC-кодов. Это также обоснованно тем, что MCC-код является кодом терминала, при этом на практике бывает так, что идентичным магазинам будут сопоставлены разные MCC-коды. При этом существует большой класс людей, посещающих лишь определенные магазины. Для статистической модели поведение людей посещающих одинаковые магазины с разными MCC-кодами различно.

Исходное множество MCC-кодов было разбито на 93 группы. Для клиента d в качестве n_{wd} считалась сумма расходов по группе w. После внесения данных изменений были повторно построены модели PLSA и LDA. На рисунке 2 представлена зависимость перплексии от номера итерации после группировки, а примеры типов потребления представлены в таблице 4. Данный прием позволил избавится от дублирующихся типов потребления.

Рис. 2. Графики перплексии для модели PLSA и LDA на группах MCC-кодов



Результаты проведённых экспериментов показали, что модели LDA и PLSA дают практически идентичный результат. Графики перплексии (1), (2) показывают незначительное преимущество модели PLSA.

3.3 Модель ARTM

Следующим шагом в построении тематической модели является регуляризация. Естественными ограничениями является наложение на матрицы Φ и Θ

Таблица 4. Интерпретируемые профили потребления для модели PLSA и LDA

PLSA		**LDA**	
φ_{wt}	**Название группы**	φ_{wt}	**Название группы**
0.47	видеоигры	0.58	косметология
0.20	услуги интернета (провайдеры и подписки)	0.14	продукты питания
0.08	магазины электроники	0.05	магазин электроники
0.05	услуги рекламные	0.04	цветочные магазины
0.05	универмаг	0.02	дом
0.03	оплата телефона	0.02	жд билеты
0.03	продукты питания	0.02	аптеки

следующих регуляризаторов: разреженности Θ, декоррелирования Φ, разреженности Φ. Это обоснованно тем, что каждый человек проявляет небольшое количество уникальных типов поведения в исследуемый период, небольшим набором MCC-кодов или соответствующих им групп.

Были построены две модели: в первой в качестве w были MCC-коды, во второй — группы MCC-кодов. Для подбора коэффициентов регуляризации использовался следующий подход в обеих моделях: из нескольких значений коэффициента выбиралось то, которое обеспечивало максимальное действие регуляризатора и при этом не увеличивало перплексию. Также для регуляризаторов сглаживания и разреживания коэффициенты α_{td}, β_{wt} были заданы константами ($\alpha_{td} = \alpha, \beta_{wt} = \beta$). Регуляризаторы добавлялись последовательно к уже обученной базовой модели PLSA в следующем порядке: разреживание предметных тем матрицы Φ, сглаживание фоновых тем матрицы Φ, декоррелирование Φ, разреживание Θ.

График зависимости перплексии от номера итерации представлен для модели без группировки MCC-кодов на рисунке 3 и с группировкой — на рисунке 4.

На графиках разреженности матрицы Φ рисунки 5, 6 видно, что количество нулевых элементов значительно увеличилось. Профили потребления теперь описываются меньшим количеством MCC-кодов, или групп. Это соответствует предположению о том, что в регуляризованной модели каждый профиль потребления будет описываться небольшим набором MCC-кодов, групп MCC-кодов.

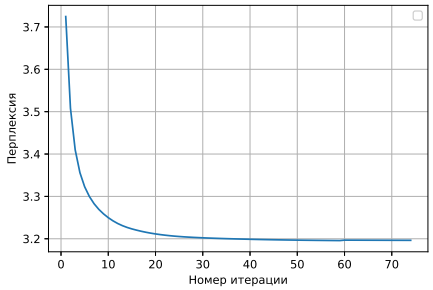


Рис. 3. Перплексия для ARTM модели на MCC-кодах

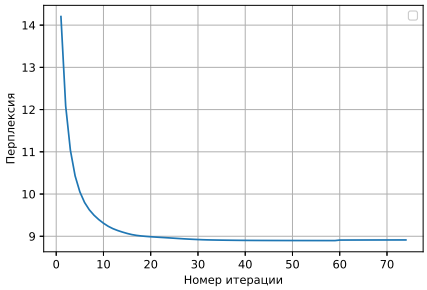


Рис. 4. Перплексия для ARTM модели на группах MCC-кодов

Декореллирование матрицы Φ приводит к уменьшению среднего стандартного коэффициента корреляции между профилями потребления φ_t с 0.041 до 0.012 для модели на группах и уменьшению с 0.22 до 0.11 для модели на MCC-кодах.

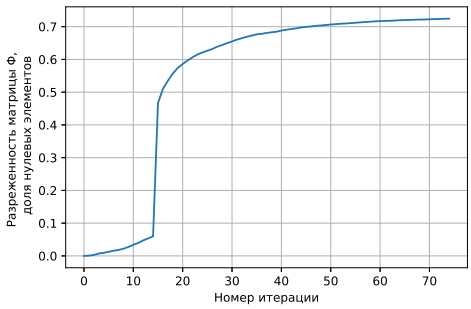


Рис. 5. Разреженность Φ для ARTM модели на MCC-кодах

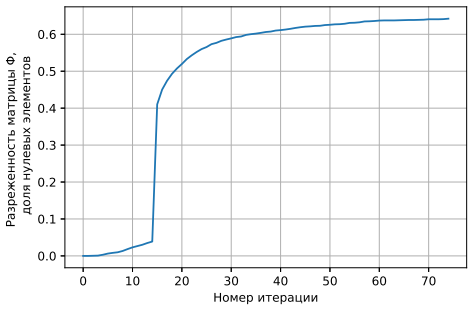


Рис. 6. Разеженность Φ для ARTM модели на группах MCC-кодов

Разреживание матрицы Θ приводит к изменению количества проявляемых каждым клиентом типов поведения. Это демонстрируют распределения количества профилей потребления у клиентов для базовой модели PLSA и регуляризованной модели ARTM представленые на рисунке 7.

В модели с меньшим словарем, построенной на группах MCC-кодов, ситуация схожая. Регуляризатор разреженности матрицы Θ также позволил сократить количество профилей потребления, которыми описывается каждый из клиентов. Соответствующие гистограммы представлены на рисунке 8.

По сравнению с базовой моделью, набор профилей модели ARTM более

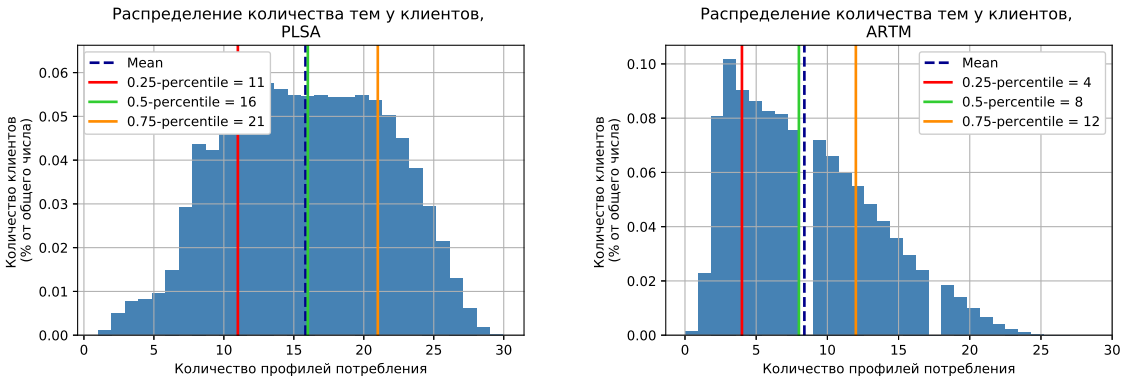


Рис. 7. Распределение количества профилей потребления в модели PLSA, ARTM, построенной на MCC-кодах

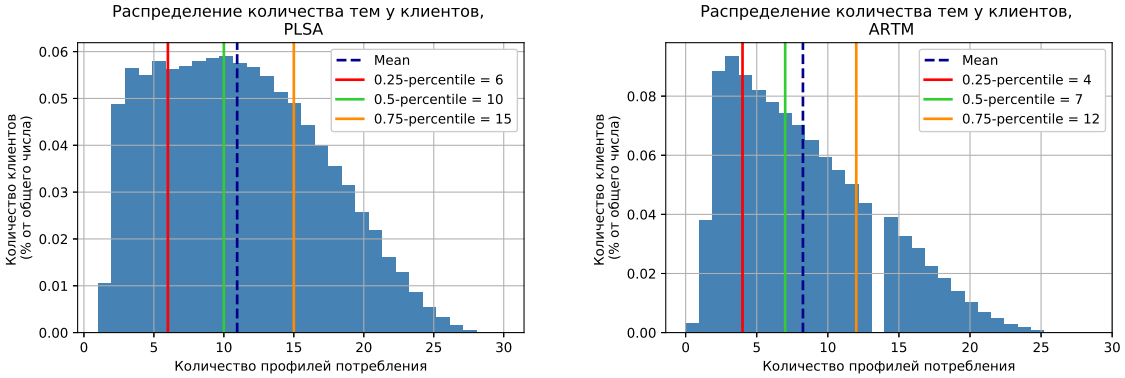


Рис. 8. Распределение количества профилей потребления в модели PLSA, ARTM, построенной на группах MCC-кодов

качественный так как: удалось уменьшить корреляции между профилями, исключить дублирование профилей; уменьшить количество MCC-кодов, которые входят в профиль потребления с ненулевым весом; повысить степень разреженности Θ, описать клиента меньшим количеством профилей потребления. Данные изменения не уменьшили правдоподобие модели, но сделали её более ценной для приложений.

3.4 Мультимодальная тематическая модель

Перейдем к модели, учитывающей дополнительные признаки пользователя, — мультимодальной тематической модели.

Модальность MCC-кодов, групп MCC-кодов в дальнейшем будем назы-

вать основной, модальности признаков — дополнительными. Модель позволяет без уменьшения правдоподобия по основной модальности сделать профили потребления согласованными с признаками клиентов. Например, тип потребления, содержащий MCC-коды магазинов косметики, салонов красоты и магазинов одежды будет согласован с женским полом, а содержащий информацию о магазинов садовых принадлежностей, магазинов инструментов — с пожилым возрастом. Весовые коэффициенты подбираются из следующих соображений: дополнительные модальности не должны увеличивать перплексию по основной модальности, среди типов потребления должны быть как типы потребления, согласованные с дополнительной модальностью, так и без однозначной корреляции с ней. Последнее предположение следует из того, что проявление некоторых типов поведения присуще людям с разным полом, возрастом, социальным положением и образованием.

Модальности добавлялись к регуляризованной модели ARTM. Использование признаков в качестве модальностей проводилось следующим образом. Все некатегориальные признаки (например, возраст) преобразовывались в категориальные. Возраст был категоризован в соответствии с возрастной стратификацией: $14-17$ — подростковый возраст, $17-25$ — юность, $25-45$ — средний возраст, >45 — пожилой возраст. В качестве слова w модальности m выбирались значения категориальных признаков, n_{dw} равно 1, если у клиента d данное значение признака, иначе 0.

Графики перплексии основной модальности модели, построенной на MCC-кодах, в точности совпадают с графиками для модели ARTM представленных на рисунках 3, 4. Это подтверждает тот факт, что добавление модальностей не ухудшило правдоподобия исходной модели. Log-правдоподобие, взятое отдельно по дополнительным модальностям, также сходится.

Основным результатом проведенного эксперимента является выделение интерпретируемых профилей потребления, согласованных с дополнительными модальностями. Такие типы потребления представлены в таблицах 5, 6.

Таким образом, применение мультимодальных тематических моделей позволяет эффективно учесть при построении профилей потребления дополнительную информацию, описывающую социальное положение клиента.

Таблица 5. Пример профиля для мультимодальной модели, построенной на MCC-кодах

φ_{wt}	**Описание признака**
Основная модальность	
0.62	Лесо-строительные материалы
0.10	Товары для дома
0.03	Садовые принадлежности (в том числе для ухода за газонами) в розницу
0.02	Различные магазины и специальные розничные магазины(карты, атласы, боеприпасы, лед, дистиллированная вода, аксессуары для магии)
0.02	Бытовое оборудование
0.02	Оборудование, мебель и бытовые принадлежности(кроме электрооборудования)
0.02	Бакалейные магазины, супермаркеты
0.01	Автозапчасти и аксессуары
0.01	Телеком: локальные и дальние телефонные звонки и услуги факса, предоплачиваемые телефонные услуги
0.01	Розничная продажа стекла, красок и обоев
Модальность пола	
0.93	Мужской
0.07	Женский
Модальность возраста	
0.33	Средний возраст
0.66	Старость
Модальность образования	
0.45	Среднее-специальное
0.31	Высшее
0.14	Среднее

3.5 Оценка предсказательной способности тематических моделей на задаче классификации

Хорошо построенный профиль потребления должен описывать человека с различных сторон. Математически оценить качество получившихся профилей потребления можно, построив алгоритм, который будет предсказывать известные признаки характеризующие человека.

Таблица 6. Пример профиля для мультимодальной модели, построенной на группах MCC-кодов

φ_{wt}	**Описание признака**	φ_{wt}	**Описание признака**
Основная модальность		**Модальность пола**	
0.43	эксклюзивная одежда	0.97	Мужской
0.31	ветеринары и зоотовары	0.03	Женский
0.08	женская одежда	**Модальность возраста**	
0.07	продукты питания	0.32	Средний возраст
0.03	аптеки	0.60	Старость
0.02	активный отдых	0.08	Юность
0.02	магазины электроники	**Модальность образования**	
		0.16	Среднее-специальное
		0.69	Высшее
		0.02	Среднее

Для достижения этой цели были построены модели логистической регрессии и градиентного бустинга, которые по построенным профилям потребления предсказывали пол и возраст. Возраст был категоризован тем же способом, что и в случае мультимодальной тематической модели. Также классификация производилась на исходных столбцах матрицы F. Данная процедура позволяет оценить качество профилей потребления в сравнение с изначальным описанием через MCC-коды, группы MCC-кодов. Результаты классификации для задачи прогнозирования пола приведены в таблице 7, возрастной категории — в таблице 8.

Результаты эксперимента свидетельствуют о двух фактах. Во-первых, переход от признаков — суммы расходов по MCC-кодам, группам MCC-кодов — к построенным типам потребления, что математически является снижением размерности признакового пространства, не привело к ухудшению результата.

Более того, мультимодальная регуляризованная тематическая модель демонстрирует наилучший результат, если в качестве исходных признаков берутся расходы по группам MCC-кодов в задаче прогнозирования пола клиента. в задаче прогнозирования возрастной категории качество всех моделей отличается незначительно. Закономерно более высокое качество модели градиентного бустинга в сравнении с логистической регрессией обусловлено тем, что размер обучающей выборки, которая во всех экспериментах бралась как 75% всех клиентов, намного превышает размер признакового пространства,

Таблица 7. Результаты классификации для задачи прогнозирования пола

Модель	**Логистическая регрессия**			**Градиентный бустинг**		
	Точность	**Полнота**	**F_1-мера**	**Точность**	**Полнота**	**F_1-мера**
Модели, построенные на MCC-кодам						
Модель на матрице N	0.73	0.72	0.72	0.74	0.71	0.72
PLSA	0.69	0.68	0.68	0.73	0.72	0.725
LDA	0.67	0.68	0.67	0.72	0.73	0.73
ARTM	0.69	0.67	0.67	0.72	0.72	0.72
Мультимодальная ARTM	0.69	0.67	0.67	0.73	0.73	0.73
Модели, построенные на группах MCC-кодов						
Модель на матрице N	0.72	0.725	0.72	0.72	0.71	0.72
PLSA	0.71	0.72	0.71	0.73	0.72	0.725
LDA	0.71	0.72	0.72	0.72	0.72	0.72
ARTM	0.72	0.71	0.71	0.74	0.74	0.74
Мультимодальная ARTM	0.73	0.74	0.74	0.75	0.75	0.75

Таблица 8. Результаты классификации для задачи прогнозирования возрастной категории

Модель	**Логистическая регрессия**			**Градиентный бустинг**		
	Точность	**Полнота**	**F_1-мера**	**Точность**	**Полнота**	**F_1-мера**
Модели, построенные на MCC-кодам						
Модель на матрице N	0.57	0.54	0.55	0.64	0.60	0.60
PLSA	0.58	0.53	0.54	0.63	0.58	0.59
LDA	0.59	0.53	0.55	0.64	0.57	0.59
ARTM	0.56	0.52	0.52	0.61	0.54	0.56
Мультимодальная ARTM	0.58	0.53	0.55	0.62	0.58	0.59
Модели, построенные на группах MCC-кодов						
Модель на матрице N	0.55	0.54	0.53	0.63	0.54	0.58
PLSA	0.58	0.51	0.53	0.60	0.55	0.57
LDA	0.59	0.51	0.53	0.63	0.57	0.60
ARTM	0.61	0.54	0.58	0.62	0.55	0.58
Мультимодальная ARTM	0.62	0.56	0.58	0.64	0.57	0.61

равный < 600. Это позволяет модели находить сложные, нелинейные закономерности в данных.

4 Заключение

Рассмотрена задача профилирования клиентов, а именно построение профиля потребления, состоящего из интерпретируемых типов поведения.

Поставленная задача решалась методом тематического моделирования. На реальных данных были построены модели PLSA, LDA. Показана их способность сходится к интерпретируемым наборам типов поведения. Также показано, что без уменьшения правдоподобия базовые модели могут быть обогащены регуляризаторами в соответствии с предметной областью. Построена мультимодальная модель, которая также без уменьшения правдоподобия по основной модальности согласует наборы типов потребления с известными признаками клиентов. Сравнение способности всех моделей на задаче предсказания пола и возрастной категории показывает, что переход к небольшому количеству интерпретируемых признаков сохраняет информацию о прогнозируемых классах, то есть профили потребления описывают клиентов также хорошо, как исходные данные.

Предполагается, что данный метод профилирования может быть применен к широкому спектру прикладных задач, таких как профилирование логов пользователей онлайн-сервисов, интернет-магазинов.

Данная задача имеет большой научный потенциал. Количество типов поведения, их несбалансированность по покрытию клиентов и по объему транзакций могут стать объектом дальнейших исследований.

Список литературы

[1] Blei David M, Ng Andrew Y, Jordan Michael I. Latent dirichlet allocation // Journal of machine Learning research. 2003. Т. 3, № Jan. С. 993–1022.

[2] Hofmann Thomas. Probabilistic latent semantic analysis // Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence / Morgan Kaufmann Publishers Inc. 1999. С. 289–296.

[3] Vorontsov Konstantin, Potapenko Anna. Additive regularization of topic models // Machine Learning. 2015. Т. 101, № 1-3. С. 303–323.

[4] Xu Guandong, Zhang Yanchun, Yi Xun. Modelling user behaviour for web recommendation using lda model. 2008. Т. 3. С. 529–532.

[5] Jin Xin, Zhou Yanzan, Mobasher Bamshad. A unified approach to personalization based on probabilistic latent semantic models of web usage and content. 2004.

[6] Mobasher Bamshad. Data mining for web personalization // The adaptive web. Springer, 2007. С. 90–135.

[7] McLachlan Geoffrey, Krishnan Thriyambakam. The EM algorithm and extensions. John Wiley & Sons, 2007. Т. 382.

[8] Bigartm: Open source library for regularized multimodal topic modeling of large collections / Konstantin Vorontsov, Oleksandr Frei, Murat Apishev [и др.] // International Conference on Analysis of Images, Social Networks and Texts / Springer. 2015. С. 370–381.

[9] Non-Bayesian additive regularization for multimodal topic modeling of large collections / Konstantin Vorontsov, Oleksandr Frei, Murat Apishev [и др.] // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications / ACM. 2015. С. 29–37.

[10] Potapenko Anna, Vorontsov Konstantin. Robust PLSA performs better than LDA // European Conference on Information Retrieval / Springer. 2013. С. 784–787.

[11] Peng Chao-Ying Joanne, Lee Kuk Lida, Ingersoll Gary M. An introduction to logistic regression analysis and reporting // The journal of educational research. 2002. T. 96, № 1. C. 3–14.

[12] Friedman Jerome H. Greedy function approximation: a gradient boosting machine // Annals of statistics. 2001. C. 1189–1232.