Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Хрыльченко Кирилл Ярославович

Обобщенные модальности в вероятностных тематических моделях для транзакционных данных

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:

д.ф-м.н., доцент

К. В. Воронцов

Москва, 2020

Содержание

1 Введение

В современном мире объем транзакционных данных растет с экспоненциальной скоростью. Это относится как к стандартным транзакционным данным, таким как покупки продуктов, билетов и других услуг; так и к *e-commerce* — транзакционной деятельности в интернете, за счет которой происходит большой рост общего объема информации.

Количество данных позволяет моделировать транзакционную деятельность — формировать профили потребителей, виды экономической деятельности, предсказывать последующие транзакции, рекомендовать продукты, находить взаимосвязь между участниками транзакций, выявлять аномальную транзакционную активность.

Мультимодальное тематическое моделирование позволяет сформировать для объектов изучения интерпретируемые векторные представления, используя при этом большое количество доступной разнородной информации, сформированной в виде *модальностей*. Например: множество контрагентов компании, множество товарных слов в платежных поручениях транзакций, в которых компания является продавцом; суммы транзакций.

Важность модальностей для модели определяется с помощью соответствующих весов, в связи с чем возникают две задачи. Первая задача — как задать веса модальностей, чтобы модель обобщила наибольшее количество информации, другими словами, учла все модальности. Вторая задача — как задать веса модальностей, чтобы сохранить в тематической модели максимальное количество релевантной информации для конкретной вспомогательной задачи. Примером такой задачи может служить предсказание просрочки по кредиту, оттока клиентов, количества полетов.

Существенным недостатком мультимодальных тематических моделей является отсутствие методики учета «вещественной информации» — таких характеристик транзакций и клиентов, как суммы транзакций, возраст клиента, длительность транзакционной жизни клиентов, количество покупок в рамках транзакции.

В данной работе формализуется понятие *вклада модальностей* в тематическую модель, на основе которого, в свою очередь, вводится понятие *равного вклада* модальностей и понятие наилучших весов модальностей для вспомогательной задачи. Кроме того, в данной работе вводятся *вещественные модальности*, позволяющие в рамках тематической модели учитывать численные характеристики. На примере задачи кредитного скоринга демонстрируется преимущество такого подхода по отношению к стандартному тематическому моделированию.

2 Основные понятия

2.1 Тематическое моделирование

Пусть $D = \{d_1, d_2, \dots\}$ — коллекция [1] сущностей, являющихся основными объектами моделирования, далее называемых **документами**. В свою очередь, документы состоят из **токенов** — в некотором смысле более «мелких» сущностей, формирующих словарь W.

Каждый документ $d \in D$ представляется как мультимножество [2] токенов $w_1, \dots, w_{n_d}$, где $w_i \in W \ \forall i$ и n_d — мощность документа [3].

Длиной коллекции $n := \sum_{d \in D} n_d$ является суммарная длина всех документов.

2.1.1 Частотные оценки

Наиболее простой вероятностной моделью для коллекции документов является частотная модель. В рамках данной работы используется термин «частотная оценка», поэтому посмотрим на данную модель более детально.

Рассмотрим *множество элементарных исходов* $\Omega = W \times D$, состоящее из всевозможных упорядоченных пар (w, d), где $w \in W$, $d \in D$. Тогда пара $(\Omega, 2^\Omega)$ образует *измеримое пространство*, где 2^Ω — класс всех подмножеств Ω [4]. Если предположить наличие *вероятностного пространства*, а именно, наличие *вероятностной меры* $\mathbb{P}$ на рассматриваемом измеримом пространстве, а также предположить, что неизвестная вероятностная мера принадлежит параметрическому семейству вероятностных мер $\mathcal{P}_\Theta = \{P_\theta, \theta \in \Theta\}$, согласованных с тем же измеримым пространством, то тройка $(\Omega, 2^\Omega, \mathcal{P}_\Theta)$ образует **статистическую структуру**.

Рассматриваемая нами коллекция документов является реализацией $(w_i, d_i)_{i=1}^n$ выборки $(\mathcal{W}_i, \mathcal{D}_i)_{i=1}^n$, элементы которой подчиняются распределению $\mathbb{P}$.

Тогда имеем следующие *оценки максимальноо правдоподобия*:

$$p(w \mid d) = \frac{n_{dw}}{n_d}, \quad p(d) = \frac{n_d}{\sum_{d \in D} n_d},$$

где n_{dw} — количество раз, которое токен w встретился в документе d.

Такие оценки на условное распределение токена в документе и вероятность документа в коллекции называются **частотными оценками**.

---

[1] В данной работе термин «коллекция» является синонимом «множества».

[2] Принимается гипотеза **мешка слов** — отсутствие чувствительности модели к порядку токенов.

[3] Длина документа.

[4] Класс всех подмножеств Ω является наиболее популярной *сигма-алгеброй* для дискретных множеств элементарных исходов.

2.1.2 PLSA

Основную гипотезу тематического моделирования, предложенную Томасом Хоффманом в модели PLSA [1], можно сформулировать следующим образом: каждой паре <*токен, документ*> (w, d), встречающейся в коллекции, соответствует **тема** t, при этом вероятность токена w в документе зависит только от распределения документа по темам.

Появляется вероятностное пространство с множеством элементарных исходов $\Omega = W \times D \times T$ и вероятностной мерой P. Токены w и документы d являются наблюдаемыми переменными, а темы t — скрытыми.

Тогда условие «вероятность токена w в документе зависит только от распределения документа по темам» можно записать в виде **гипотезы условной независимости**:

$$p(w \,|\, t, d) = p(w \,|\, t). \tag{1}$$

Распределение токена $w \in W$ при условии документа $d \in D$ принимает вид:

$$p(w \,|\, d) = \sum_{t \in T} p(w \,|\, t, d) p(t \,|\, d) = \{1\} = \sum_{t \in T} p(w \,|\, t) p(t \,|\, d) = \sum_{t \in T} \phi_{wt} \theta_{td},$$

где $\Phi \in \mathbb{R}^{|W| \times |T|}, \Theta \in \mathbb{R}^{|T| \times |D|}$ — стохастические матрицы[5], которые являются параметрами тематической модели.

Оптимальными значениями параметров Φ, Θ считается решение задачи максимизации логарифма правдоподобия наблюдаемой коллекции документов:

$$\begin{cases} \sum\limits_{d \in D} \sum\limits_{w \in W} n_{dw} \log p(w \,|\, d, \Phi, \Theta) = \sum\limits_{d \in D} \sum\limits_{w \in W} n_{dw} \log \sum\limits_{t \in T} \phi_{wt} \theta_{td} \to \max\limits_{\Phi, \Theta}, \\ \Phi, \Theta \text{ — стохастические матрицы.} \end{cases}$$

Данную задачу оптимизации можно решать с помощью EM-алгоритма [3] — итеративного метода, состоящего из двух шагов:

E-шаг. С помощью теоремы Байеса оценивается распределение скрытой переменной $p(t \,|\, d, w, \Phi, \Theta)$ для всех токенов $w \in W$ и документов $d \in D$:

$$p_{tdw} = p(t \,|\, d, w, \Phi, \Theta) = \big\{\text{теор. Байеса}\big\} = \frac{p(w \,|\, t, \Phi) p(t \,|\, d, \Theta)}{p(w \,|\, d, \Phi, \Theta)} = \frac{\phi_{wt} \theta_{td}}{\sum\limits_{s \in T} \phi_{ws} \theta_{sd}}.$$

M-шаг является решением оптимизационной задачи $\mathbb{E}_{p_{tdw}} \log p(w, d, t | \Phi, \Theta) \to \max_{\Phi, \Theta}$, которое можно получить с помощью теоремы Каруша-Куна-Таккера:

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \right), \quad \theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \right),$$

где $\operatorname{norm}_{x \in X} f(x) = \frac{f(x)}{\sum_{x \in X} f(x)}$.

[5]Матрица $F \in \mathbb{R}^{m \times n}$ является стохастической, если $F_{ij} \geqslant 0$ и $\sum_{i=1}^{m} F_{ij} = 1$, то есть столбцы образуют вероятностные распределения.

2.1.3 Мультимодальное тематическое моделирование

Рассмотрим ситуацию, в которой документы состоят из токенов разной природы. Например, статья состоит из текста и списка цитируемой литературы. Логично предположить, что токены разной природы не лежат в одном распределении по темам. Каждому источнику данных m, называемому *модальностью*, можно сопоставить свой словарь $W_m \in M$, где M — множество модальностей, и распределение $p(w \mid t)$, $w \in W_m$, информация о котором формируется в виде матрицы $\Phi_m = (\phi_{wt} := p(w \mid t))$.

По аналогии с PLSA, максимизируется взвешенная сумма логарифмов правдоподобия модальностей:

$$\begin{cases} \sum\limits_{m \in M} \lambda_m \sum\limits_{d \in D} \sum\limits_{w \in W_m} n_{dw} \log \sum\limits_{t \in T} \phi_{wt} \theta_{td} \to \max\limits_{\Phi, \Theta}, \\ \Phi_m \ \forall m, \ \Theta \text{ — стохастические матрицы}, \end{cases} \tag{2}$$

где $\lambda_m \geqslant 0$ — веса модальностей. Для удобства обозначений формируется матрица Φ, образованная конкатенацией записанных в столбец матриц Φ_m, при этом предполагается, что словари различных модальностей не пересекаются — $W_i \cap W_j = \emptyset$ при $i \neq j$.

EM-алгоритм для мультимодальных тематических моделей выводится аналогично PLSA, при этом E-шаг и обновление матрицы Φ не меняются, а вычисление тематических векторных представлений для документов на M-шаге принимает вид:

$$\theta_{td} = \operatorname{norm}_t \left(\sum_{m \in M} \lambda_m \sum_{w \in W} n_{dw} p_{tdw} \right).$$

2.1.4 Рациональный EM-алгоритм для тематического моделирования

Алгоритм 1 позволяет обучать тематическую модель без необходимости хранить p_{tdw} в памяти целиком, итеративно вычисляя значение p_{tdw} только для документа d и слова w.

Алгоритм 1 Рациональный EM-алгоритм для тематической модели.

Вход: коллекция D, количество тем $|T|$, начальные приближения матриц Φ и Θ

Выход: параметры модели Φ и Θ

1: **повторять**
2: обнулить n_{wt}, n_{td}, n_t для всех $d \in D,\ w \in W,\ t \in T$
3: **для всех** $d \in D,\ w \in d$
4: $n_{tdw} := n_{dw} \phi_{wt} \theta_{td} / \sum_\tau \phi_{w\tau} \theta_{\tau d}$ для всех $t \in T$;
5: увеличить n_{wt}, n_{td}, n_t на n_{tdw} для всех $t \in T$;
6: $\phi_{wt} := n_{wt} / n_t$ для всех $w \in W, t \in T$;
7: $\theta_{td} := n_{td} / n_d$ для всех $d \in D, t \in T$;
8: **пока** Φ и Θ не сойдутся

3 Взвешивание модальностей

Стандартной практикой задания весов модальностей является присваивание всем весам единичных значений, $\lambda_m = 1 \; \forall m$, с последующим «ручным подбором» весов путем умножения или деления определенных весов на небольшие по модулю положительные числа методом «пристального взгляда». Конечной целью таких операций является либо получение наиболее интерпретируемых результатов, либо оптимизация вспомогательного критерия.

Возникает проблема нечувствительности модели к таким изменениям: как далее в работе будет показано, если в среднем количество токенов одной модальности в документе много больше, чем количество токенов другой модальности, то первая модальность будет «заглушать» влияние другой модальности на тематические представления. Поэтому существует необходимость разработки теоретически обоснованной методики выравнивания модальностей, учитывающей всю имеющуюся информацию.

«Уравнивающие» веса модальностей, которые вводятся в данной работе, являются хорошим начальным приближением для дальнейших изменений соотношения модальностей: увеличение тех или иных весов в n раз действительно будут увеличивать важность в n раз по отношению к другим модальностям.

Как правило, тематическое моделирование используется для решения конкретной задачи, которую можно сформулировать как оптимизационную задачу для некоторого вспомогательного критерия. В такой ситуации необходимо подбирать веса под задачу. В курсовой работе [4], автор предпринял успешную попытку автоматического подбора весов модальностей. Алгоритм состоял из жадного пошагового добавления модальностей, веса которых определялись в виде выпуклой комбинации. На каждой итерации подбиралось соотношение весов всех предыдущих модальностей с новой модальностью с помощью метода золотого сечения. С помощью такого подхода, на транзакционных данных корпоративных клиентов Сбербанка удалось получить значительное улучшение показателей качества решения задачи определения сходства экономической деятельности компаний.

Главным минусом такой методики является необходимость обучения новой тематической модели на каждой итерации алгоритма, а также сомнительная оптимальность получаемого решения из-за жадного характера оптимизации. В данной работе предлагается методика автоподбора весов модальностей для оптимизации вспомогальных критериев, подбирающая веса во время EM-алгоритма без необходимости в повторных обучениях тематической модели. Кроме того, предлагаемая методика не несет жадный пошаговый характер и ищет оптимальную комбинацию весов в $\lambda \in \mathbb{R}^m$.

3.1 Мультимодальные разложения

Приводимые ниже результаты являются основополагающими для данной работы и позволяют улучшить качество любой мультимодальной тематической модели с помощью уравнивания весов модальностей, а также являются удобным инструментом для наглядного исследования влияния мультимодального обучения на отдельные модальности. Кроме того, данные результаты лежат в основе предлагаемой далее техники автоподбора весов модальностей для оптимизации вспомогательных критериев.

Зафиксируем условное распределение скрытой переменной p_{tdw}, другими словами, сделаем E-шаг EM-алгоритма. Обозначим как n_d^m мощность модальности m в документе d. Тогда для мультимодальной тематической модели на M-шаге справедлива следующая цепочка тождеств:

$$\begin{aligned}
\theta_{td} &= \frac{\sum_{m\in M}\lambda_m\sum_{w\in W_m} n_{dw}p_{tdw}}{\sum_{t\in T}\sum_{m\in M}\lambda_m\sum_{w\in W_m} n_{dw}p_{tdw}} = \frac{\sum_{m\in M}\lambda_m\sum_{w\in W_m} n_{dw}p_{tdw}}{\sum_{m\in M}\lambda_m\sum_{w\in W_m} n_{dw}\sum_{t\in T} p_{tdw}} = \\
&= \left\{\sum_{t\in T} p_{tdw} = 1\right\} = \frac{\sum_{m\in M}\lambda_m\sum_{w\in W_m} n_{dw}p_{tdw}}{\sum_{m\in M}\lambda_m\sum_{w\in W_m} n_{dw}} = \left\{\sum_{w\in W_m} n_{dw} = n_d^m\right\} = \\
&= \frac{\sum_{m\in M}\lambda_m\sum_{w\in W_m} n_{dw}p_{tdw}}{\sum_{m\in M}\lambda_m n_d^m} = \sum_{m\in M}\frac{\lambda_m n_d^m}{\sum_{m\in M}\lambda_m n_d^m}\frac{\sum_{w\in W_m} n_{dw}p_{tdw}}{n_d^m} = \\
&= \sum_{m\in M}\tau_d^m\theta_{td}^m, \qquad \text{где } \tau_d^m = \frac{\lambda_m n_d^m}{\sum_{m\in M}\lambda_m n_d^m}, \quad \theta_{td}^m = \frac{\sum_{w\in W_m} n_{dw}p_{tdw}}{n_d^m}.
\end{aligned}$$

Будем называть τ_d^m важностью модальности m для документа d, а θ_{td}^m — **мономодальным** тематическим векторным представлением для документа d и модальности m. Для них справедливы следующие утверждения:

- $\sum_{m\in M}\tau_d^m = 1$ и $\tau_d^m \geqslant 0$, то есть τ_d^m является симплексом и образует распределение по модальностям
- веса $\lambda = \hat{1} \in \mathbb{R}^M$ имеют следующую вероятностную интерпретацию: вероятность получить токен модальности m при случайном выборе токена из документа d равна τ_d^m
- аналогично для $\lambda_m = (n_d^m)^{-1}$: при случайном выборе токена из документа d вероятности разных модальностей равны
- если $|M| = 1$, то $\theta_{td}^m = \theta_{td}$
- Мономодальные тематические векторные представления равны тематическим представлениям, получаемым при обучениии M тематических моделей с одной модальностью с одинаковой инциализацией. Такое обучение мы далее будем называть **мономодальным**

А значит, **мультимодальное** тематическое векторное представление для документа d, получаемое с помощью мультимодальной тематической модели, представляется в виде выпуклой комбинации **мономодальных** тематических векторных представлений для документа d, которые можно получить с помощью раздельного M-шага по всем модальностям. Вес, с которым мономодальное представление модальности $m \in M$ участвует в выпуклой комбинации, прямопропорционален мощности модальности в документе, равной n_d^m, а также весу модальности λ_m. Данный результат можно сформулировать в виде теоремы:

Теорема 1 (О мультимодальном разложении Θ) *Вероятность θ_{td} темы t в документе d можно представить в виде выпуклой комбинации элементов θ_{td}^m:*

$$\theta_{td} = \sum_{m=1}^{M} \tau_d^m \theta_{td}^m \quad \forall\, t \in T,\, d \in D, \qquad \text{где } \tau_d^m = \frac{\lambda_m n_d^m}{\sum_{m=1}^{M} \lambda_m n_d^m},\ \theta_{td}^m = \frac{\sum_{w \in W_m} n_{dw} p_{tdw}}{n_d^m}. \tag{3}$$

△ Доказательство приведено выше. □

Следствие 1 *Для одного шага EM-алгоритма равный вклад модальностей $m \in M$ в тематическое векторное представление документа $d \in D$ обеспечивается весами модальностей $\lambda_m = (n_d^m)^{-1}$.*

△ Равный вклад модальностей эквивалентен заданию равных весов мономодальных тематических представлений в выпуклой комбинации мультимодального векторного представления:

$$\theta_{td} = \sum_{m=1}^{M} \tau_d^m \theta_{td}^m = \sum_{m=1}^{M} \frac{\lambda_m n_d^m}{\sum_{m \in M} \lambda_m n_d^m} \theta_{td}^m = \left\{\lambda_m = (n_d^m)^{-1}\right\} = \sum_{m=1}^{M} \frac{1}{|M|} \theta_{td}^m.$$

□

При таком выборе весов, все модальности будут вносить одинаковый вклад в итоговое мультимодальное тематическое представление документа.

Для документов с разными мощностями модальностей постановка задачи (4) не позволяет уравновесить модальности для всех документов сразу. Предлагается использовать следующее приближение весов модальностей:

$$\lambda_m = \mathbb{E}_{p(d)} (n_d^m)^{-1} = \sum_{d \in D} \frac{n_d}{\sum_{d \in D} n_d} (n_d^m)^{-1},$$

где $n_d = \sum_{m \in M} n_d^m$ — общая длина документа d, определяемая как суммарная длина документа по всем модальностям.

Другим способом уравнивания модальностей является переход к постановке задачи, подразумевающей взвешивание документов. Основным минусом такого подхода, как будет показано в следующей секции, является возникающая нормировка документов — веса, уравнивающие модальности, также уравнивают вклад документов в тематические представления матрицы Φ.

Теорема 2 (О мультимодальном разложении p_{tdw}) *Условное распределение p_{tdw} темы t при условии документа d и слова w можно представить в виде выпуклой комбинации элементов p^m_{tdw}:*

$$p_{tdw} = \sum_{m=1}^{M} \eta_m^{dw} p_{tdw}^m \quad \forall\, t \in T,\, d \in D,\, w \in \cup_{m=1}^{M} W_m,$$

где

$$\eta_m^{dw} = \frac{\lambda_m n_d^m \sum_{t\in T} \phi_{wt}\theta_{td}^m}{\sum_{m\in M} \lambda_m n_d^m \sum_{t\in T} \phi_{wt}\theta_{td}^m} = \tau_d^m \frac{p(w|d,\Theta_m)}{p(w|d,\Theta)}, \quad p_{tdw}^m = \frac{\phi_{wt}\theta_{td}^m}{\sum_{t\in T}\phi_{wt}\theta_{td}^m}.$$

△ Для доказательства необходимо воспользоваться мультимодальным разложением Θ:

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{t\in T}\phi_{wt}\theta_{td}} = \{(3)\} = \frac{\phi_{wt}\sum_{m\in M}\tau_d^m\theta_{td}^m}{\sum_{t\in T}\phi_{wt}\sum_{m\in M}\tau_d^m\theta_{td}^m} = \frac{\sum_{m\in M}\tau_d^m\phi_{wt}\theta_{td}^m}{\sum_{m\in M}\tau_d^m\sum_{t\in T}\phi_{wt}\theta_{td}^m} =$$
$$= \left\{\phi_{wt}\theta_{td}^m = p_{tdw}^m \sum_{t\in T}\phi_{wt}\theta_{td}^m\right\} = \sum_{m\in M}\frac{\tau_d^m\sum_{t\in T}\phi_{wt}\theta_{td}^m}{\sum_{m\in M}\tau_d^m\sum_{t\in T}\phi_{wt}\theta_{td}^m} p_{tdw}^m$$

□

Для мультимодального разложения p_{tdw} справедливы утверждения, аналогичные утверждениям для мультимодального разложения θ_{td}. Изучение прикладной пользы данной теоремы выходит за рамки данной работы; она не используется в предлагаемых модификациях для улучшения мультимодального тематического моделирования.

3.2 Взвешивание документов

По теореме о мультимодальном разложении 3, для обеспечения равного вклада модальностей в веса документов необходимо выбирать разные веса модальностей для документов. Поэтому необходимо рассмотреть более постановку задачи тематического моделирования с весами, различными по документам:

$$\begin{cases} \sum\limits_{m\in M}\sum\limits_{d\in D}\lambda_d^m \sum\limits_{w\in W_m} n_{dw}\log\sum_{t\in T}\phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}, \\ \Phi_m,\ m\in M;\ \Theta \text{ — стохастические матрицы.} \end{cases} \tag{4}$$

E-шаг EM-алгоритма не меняется, а обновление матрицы Φ принимает вид:

$$\phi_{wt} = \operatorname{norm}_w\left(\sum_{d\in D}\lambda_d^m n_{dw}p_{tdw}\right), \quad \theta_{td} = \operatorname{norm}_t\left(\sum_{m\in M}\lambda_d^m\sum_{w\in W} n_{dw}p_{tdw}\right).$$

Согласно теореме о мультимодальном разложении Θ:

$$\theta_{td} = \sum_{m\in M}\frac{\lambda_d^m n_d^m}{\sum_{m\in M}\lambda_d^m n_d^m}\theta_{td}^m.$$

Чтобы уравновесить вклад модальностей в тематические представления документов, веса модальностей должны быть равны $\lambda_d^m = c_d(n_d^m)^{-1}$, где c_d — произвольная подокументная константа. Необходимо отдельно исследовать влияние весов λ_d^m и, в частности, предлагаемой схемы на обновление параметров матрицы Φ во время M-шага.

Рассмотрим коллекцию $D = \{d\}$, состоящую из одного документа. Тогда вычисление матрицы Φ на M-шаге EM-алгоритма принимает вид:

$$\phi_{wt}^d := \frac{n_{dw}p_{tdw}}{\sum_{w\in W_m} n_{dw}p_{tdw}}, \quad w \in W_m.$$

Теорема 3 (О подокументном разложении Φ) *Вероятность токена w модальности m при условии темы t можно представить в виде выпуклой комбинации элементов ϕ_{wt}^d:*

$$\phi_{wt} = \sum_{d\in D} \xi_{dw}\phi_{wt}^d \quad \forall\, t \in T,\, w \in \cup_{m=1}^M W_m,$$

где

$$\xi_{dw} = norm_d\left(\lambda_d^m \sum_{w\in W_m} n_{dw}p_{tdw}\right), \quad \phi_{wt}^d = \frac{n_{dw}p_{tdw}}{\sum_{w\in W_m} n_{dw}p_{tdw}}, \quad w \in W_m.$$

△ Справедлива следующая цепочка тождеств:

$$\phi_{wt} = \frac{\sum_{d\in D}\lambda_d^m n_{dw}p_{tdw}}{\sum_{w\in W_m}\sum_{d\in D}\lambda_d^m n_{dw}p_{tdw}} = \left\{ n_{dw}p_{tdw} = \phi_{wt}^d \sum_{w\in W_m} n_{dw}p_{tdw} \right\} =$$
$$= \sum_{d\in D} \frac{\lambda_d^m \sum_{w\in W_m} n_{dw}p_{tdw}}{\sum_{d\in D}\lambda_d^m \sum_{w\in W_m} n_{dw}p_{tdw}}\phi_{wt}^d = \sum_{d\in D}\xi_{dw}\phi_{wt}^d.$$

□

Вклад документа d в распределения токенов w при условии темы t прямопропорционален $\sum_{w\in W_m} n_{dw}p_{tdw}$. Если Φ, Θ находятся в окрестности оптимума, то

$$p(w \mid d, \Phi, \Theta) \approx p(w|d) = \frac{n_{dw}}{n_d^m},$$

где $p(w|d)$ — частотная оценка вероятности токена модальности m в документе. Тогда:

$$\sum_{w\in W_m} n_{dw}p_{tdw} = \sum_{w\in W_m} n_{dw}\frac{\phi_{wt}\theta_{td}}{p(w \mid d, \Phi, \Theta)} \approx \left(\sum_{w\in W_m} \frac{n_{dw}\phi_{wt}n_d^m}{n_{dw}}\right)\theta_{td} = n_d^m\theta_{td}.$$

А значит, в окрестности оптимального решения вклад документа $d \in D$ в ϕ_{wt} прямопропорционален $n_d^m\theta_{td}$:

$$\phi_{wt} \approx \sum_{d\in D} \frac{\lambda_d^m n_d^m \theta_{td}}{\sum_{d\in D}\lambda_d^m n_d^m \theta_{td}}\phi_{wt}^d.$$

Алгоритм 2 Один шаг EM-алгоритма через мультимодальное разложение.

Вход: матрицы Φ_k и Θ_k

Выход: матрицы Φ_{k+1} и Θ_{k+1}

1: **E-шаг:** вычислить p_{tdw} для всех $t \in T, d \in D, w \in W$

2: **для всех** $m \in M$

3: вычислить Φ^m и Θ^m

4: вычислить Θ_{k+1} как выпуклую комбинацию Θ^m

5: вычислить Φ_{k+1} как конкатенацию $\Phi^1, \ldots, \Phi^m$

Следствие 2 *Для одного шага EM-алгоритма в окрестности оптимальноо решения равный вклад документов в условное распределение $\phi_{wt}, w \in W_m$ обеспечивается весами модальностей $\lambda_d^m = (n_d^m)^{-1}$.*

Таким образом, при обеспечении равного вклада модальностей также происходит уравнивание весов документов.

3.3 Оптимизация вспомогательных критериев

Уравнивание вклада модальностей позволяет получить наиболее информативные тематические векторные представления, но особый интерес представляет подбор весов модальностей, обеспечивающих наилучшее качество решения вспомогательной задачи, использующей тематические представления в качестве признакового пространства.

Алгоритм 2 демонстрирует обучение тематической модели с помощью мультимодального разложения, эквивалентное стандартной схеме обучения тематической модели. Применение данной схемы на практике более затратно по памяти, так как необходимо хранить вместо одной матрицы $|M|$ различных матриц Θ^m размерности $|T| \times |D|$. Однако появляется возможность напрямую оптимизировать веса модальностей $\lambda_m, m \in M$ для решения вспомогательных задач.

Пусть имеется вспомогательный критерий $\mathcal{J}(\Theta)$, для которого поставлена задача оптимизации. Тогда веса модальностей для оптимизации вспомогательного критерия можно определить, решая задачу:

$$\begin{cases} J(\Theta(\lambda)) \to \min_\lambda, \\ \lambda \succeq 0, \end{cases} \tag{5}$$

где $(\Theta)_{td} = \theta_{td}(\lambda) = \sum_{m=1}^{M} \frac{\lambda_m n_d^m}{\sum_{m \in M} \lambda_m n_d^m} \theta_{td}^m$ — мультимодальное разложение, а $\lambda = (\lambda_1, \ldots, \lambda_M)$ — веса модальностей.

Решение задачи бинарной классификации. Рассмотрим задачу бинарной классификации на множестве документов. В качестве признакового пространства будем использовать тематиче-

ские представления документов, полученные с помощью тематического моделирования. Тогда оценку вероятности принадлежности документа к положительному классу можно сформулировать следующим образом:

$$\hat{y}_d = \sum_{t \in T} w_t \theta_{td}, \qquad \text{где } \sum_{t \in T} w_t = 1,\ w_t \geqslant 0.$$

Теорема о мультимодальном разложении θ_{td} позволяет разложить данную оценку как выпуклую комбинацию оценок:

$$\hat{y}_d = \sum_{t \in T} w_t \sum_{m \in M} \frac{\lambda_m n_d^m}{\sum_{m \in M} \lambda_m n_d^m} \theta_{td}^m = \sum_{m \in M} \frac{\lambda_m n_d^m}{\sum_{m \in M} \lambda_m n_d^m} \hat{y}_d^m,$$

где $\hat{y}_d^m = \sum_{t \in T} w_t \theta_{td}^m$ — оценка вероятности принадлежности документа к положительному классу при использовании в качестве признакового пространства мономодальных тематических представлений документов для модальности m.

Уравнивающие веса модальностей $\lambda_m = (n_d^m)^{-1}$ приводят к усреднению предсказаний, полученных использовании всех мономодальных тематических представлений. При отсутствии дополнительной информации о модальностях такие веса действительно являются оптимальными.

Решение задачи классификации можно получить минимизацией кросс-энтропии:

$$\begin{cases} \sum\limits_{d \in D} [y_d \log \hat{y}_d + (1 - y_d) \log \hat{y}_d] \to \max\limits_{\lambda, w}, \\ \sum\limits_{t \in T} w_t = 1,\ w \succeq 0, \\ \lambda \succeq 0. \end{cases}$$

Предлагается переход к следующим параметрам: $\lambda = e^{\bar{\lambda}}$ и $w = \mathrm{softmax}(\bar{w})$. Тогда задачу можно решать методом градиентного спуска. В данной работе поочередно применяется несколько итераций градиентного спуска по w и по λ.

4 Вещественные модальности

При работе с мультимодальными тематическими моделями отсутствует возможность учета информации о документе, представленной в виде множества вещественных характеристик. Например: суммы транзакций фирмы, суммы покупок потребителя. Ниже выводится методика, позволяющая добавить в тематическую модель вещественную информацию.

Пусть каждому документу d соответствует мультимножество вещественных характеристик $x_1, \ldots, x_{n_d^m}$. Тогда можно сопоставить каждой теме гауссиану с параметрами μ_t, σ_t^2 и рассмотреть вероятностную модель, в которой мультимножество вещественных характеристик порождается смесью гауссиан.

Имеется множество модальностей $M = M_s \cup M_g$, где M_s — стандартные модальности для тематического моделирования, а M_g — вещественные модальности, которые каждой теме сопоставляют нормальное распределение. Для упрощения записей удобно принять следующие обозначения:

- $\phi_{xt}^m := \mathcal{N}(x \,|\, \mu_{mt}, \sigma_{mt}^2)$, где μ_{mt}, σ_{mt}^2 — параметры гауссианы, соответствующей теме t и модальности m
- Для вещественных модальностей множество значений совпадает с $\mathbb{R}$:

$$\sum_{x \in \mathbb{R}} n_{dx} a_{dx} := \sum_{x \in \mathbb{R}:\, n_{dx} > 0} n_{dx} a_{dx}.$$

  Суммирование по $x \in \mathbb{R}$ является конечной операцией, потому что множество $\{x \in \mathbb{R} : n_{dx} > 0\}$ конечно

Теорема 4 *EM-алгоритм для обобщенной мультимодальной тематической модели с модальностями $M = M_s \cup M_g$ имеет вид:*

$$\begin{aligned}
p_{tdw} &= norm_{t \in T}\left(\phi_{wt}\theta_{td}\right), \quad t \in T,\ d \in D,\ w \in W_m,\ m \in M_s;\\
p_{mtdx} &= norm_{t \in T}\left(\phi_{xt}^m\theta_{td}\right), \quad t \in T,\ d \in D,\ x \in \mathbb{R},\ m \in M_g;\\
\theta_{td} &= norm_{t \in T}\left[\sum_{m \in M_s}\lambda_m \sum_{w \in W_m} n_{dw}p_{tdw} + \sum_{m \in M_g}\lambda_m \sum_{x \in \mathbb{R}} n_{dx}p_{mtdx}\right], \quad t \in T,\ d \in D;\\
\phi_{wt} &= norm_{w \in W_m}\left[\sum_{d \in D} n_{dw}p_{tdw}\right], \quad w \in W_m,\ t \in T,\ m \in M_s;\\
\mu_{mt} &= \sum_{d \in D}\sum_{x \in \mathbb{R}} \frac{n_{dx}p_{mtdx}}{\sum_{d \in D}\sum_{x \in \mathbb{R}} n_{dx}p_{mtdx}} \cdot x, \quad t \in T,\ m \in M_g;\\
\sigma_{mt}^2 &= \sum_{d \in D}\sum_{x \in \mathbb{R}} \frac{n_{dx}p_{mtdx}}{\sum_{d \in D}\sum_{x \in \mathbb{R}} n_{dx}p_{mtdx}} \cdot (x - \mu_t)^2, \quad t \in T,\ m \in M_g.
\end{aligned}$$

$\triangle$ E-шаг для стандартный модальностей не меняется по сравнению с обычным мультимодальным тематическим моделированием, а для вещественных модальностей принимает вид:

$$p_{mtdx} = p(t \,|\, d, x, \mu_m, \sigma_m^2, \Theta) = \frac{p(x \,|\, t, \mu_m, \sigma_m^2)p(t \,|\, d, \Theta)}{p(x \,|\, d, \mu_m, \sigma_m^2, \Theta)} = \frac{\mathcal{N}(x \,|\, \mu_{mt}, \sigma_{mt}^2)\theta_{td}}{\sum\limits_{s \in T}\mathcal{N}(x \,|\, \mu_{ms}, \sigma_{ms}^2)\theta_{sd}} =$$

$$= \mathrm{norm}_{t \in T}\left(\phi_{xt}^m\theta_{td}\right), \quad t \in T,\ d \in D,\ x \in \mathbb{R},\ m \in M_g.$$

Примем обозначения:

$$\begin{aligned}
\mathcal{L}_s(\Phi, \Theta) &:= \sum_{m \in M_s}\lambda_m \sum_{t \in T}\sum_{d \in D}\sum_{w \in W_m} n_{dw}p_{tdw}\,\log \phi_{wt}\theta_{td};\\
\mathcal{L}_g(\mu, \sigma, \Theta) &:= \sum_{m \in M_g}\lambda_m \sum_{t \in T}\sum_{d \in D}\sum_{x \in \mathbb{R}} n_{dx}p_{mtdx}\log \mathcal{N}(x \,|\, \mu_{mt}, \sigma_{mt}^2)\,\theta_{td}.
\end{aligned}$$

Тогда M-шаг выглядит следующим образом:

$$\begin{cases} \mathcal{L}_s(\Phi,\Theta) + \mathcal{L}_g(\mu,\sigma,\Theta) \to \max\limits_{\Phi,\Theta,\mu,\sigma}, \\ \Theta, \{\Phi_m \,|\, m \in M_s\} \text{ — стохастические матрицы}, \\ \sigma \succeq 0. \end{cases} \tag{6}$$

и разбивается на несколько независимых задач оптимизации.

1. Оптимизационная задача для нахождения тематических представлений документов:

$$\begin{cases} \sum\limits_{t\in T}\left[\sum\limits_{m\in M_s}\lambda_m\sum\limits_{w\in W_m} n_{dw}p_{tdw} + \sum\limits_{m\in M_g}\lambda_m\sum\limits_{x\in\mathbb{R}} n_{dx}p_{mtdx}\right]\log\theta_{td} \to \max\limits_{\Theta}, \\ \sum\limits_{t\in T}\theta_{td} = 1, \qquad d\in D; \\ \theta_{td} \geqslant 0, \quad t\in T. \end{cases} \tag{7}$$

Решение задачи 7 целиком совпадает с решением аналогичной задачи для обычной мультимодальной тематической модели и выводится с помощью теоремы Каруша-Куна-Такера:

$$\theta_{td} = \operatorname{norm}_{t\in T}\left[\sum_{m\in M_s}\lambda_m\sum_{w\in W_m} n_{dw}p_{tdw} + \sum_{m\in M_g}\lambda_m\sum_{x\in\mathbb{R}} n_{dx}p_{mtdx}\right], \quad t\in T,\ d\in D.$$

2. Задача для определения значений матрицы Φ:

$$\begin{cases} \sum\limits_{w\in W_m}\left[\sum_{d\in D} n_{dw}p_{tdw}\right]\log\phi_{wt} \to \max\limits_{\Phi^m}, \\ \sum_{w\in W_m}\phi_{wt} = 1, \qquad m\in M_s,\ t\in T \\ \phi_{wt} \geqslant 0, \quad w\in W_m. \end{cases} \tag{8}$$

Решение задачи 8 целиком совпадает с задачей определения значений матрицы Φ на M-шаге EM-алгоритма для мультимодальной тематической модели:

$$\phi_{wt} = \operatorname{norm}_{w\in W_m}\left[\sum_{d\in D} n_{dw}p_{tdw}\right], \quad w\in W_m,\ t\in T,\ m\in M_s.$$

3. Задача определения параметров гауссиан μ, σ^2:

$$\begin{cases} \sum\limits_{d\in D}\sum\limits_{x\in\mathbb{R}} n_{dx}p_{mtdx}\left(-\log\sigma_t - (x-\mu_{mt})^2/(2\sigma_{mt}^2)\right) \to \max\limits_{\mu_{mt},\sigma_{mt}}, \qquad m\in M_g,\ t\in T \\ \sigma_{mt} > 0. \end{cases} \tag{9}$$

Для решения задачи 9 достаточно приравнять производные по параметрам к нулю и в явном виде вывести зависимость:

$$\mu_{mt} = \sum_{d\in D}\sum_{x\in\mathbb{R}} \frac{n_{dx}p_{mtdx}}{\sum_{d\in D}\sum_{x\in\mathbb{R}} n_{dx}p_{mtdx}} \cdot x, \quad t\in T,\ m\in M_g; \tag{10}$$

$$\sigma_{mt}^2 = \sum_{d\in D}\sum_{x\in\mathbb{R}} \frac{n_{dx}p_{mtdx}}{\sum_{d\in D}\sum_{x\in\mathbb{R}} n_{dx}p_{mtdx}} \cdot (x-\mu_t)^2, \quad t\in T,\ m\in M_g; \tag{11}$$

□

5 Эксперименты

Эксперименты проводятся на транзакционных данных корпоративных клиентов Росбанка с 01.01.2015 по 31.01.2019.

5.1 Показатели качества

Решается задача бинарной классификации:

- объекты классификации - состояния клиентов на все моменты времени с начала транзакционной активности до последнего периода отчетности; 83898 периодов отчетности для 5352 различных клиентов
- целевая переменная - факт просрочки 90+ дней на горизонте года; положительный класс составляет 0.063 часть выборки
- признаковое пространство - тематические векторные представления из матрицы Θ, полученной с помощью тематического моделирования
  - **стандартные модальности:**
    - товарные слова из платежных поручений
    - множество контрагентов
    - сегмент бизнеса клиента
  - **вещественные модальности:**
    - суммы транзакций

Метрика классификации. Для измерения качества решения вспомогательной задачи классификации используется метрика ROC-AUC, равная доле правильно упорядоченных алгоритмом по значению целевой переменной пар объектов.

Модель. В качестве модели классификации используется реализация градиентного бустинга [6] в библиотеке **lightgbm** [5] со значениями гиперпараметров, указанными в таблице 1. Значения, отсутствующие в данной таблице, взяты по умолчанию.

Параметр	Значение
Шаг обучения	0.1
Количество листьев	3
Количество итераций до ранней остановки	50
Метрика для ранней остановки	ROC-AUC
Максимальное количество итераций	10000

Таблица 1: Значения гиперпараметров градиентного бустинга.

Валидация. Для обеспечения возможности использования *ранней остановки* градиентного бустинга без утечки данных используется *двойная кросс-валидация*: данные для обучения делятся на n равных частей, называемых *фолдами*, и качество измеряется как средний ROC-AUC по этим «внешним» фолдам, в то время как внутри выборки, используемой для обучения модели во время одной итерации кросс-валидации, происходит еще одна кросс-валидация, в рамках которой и осуществляется ранняя остановка за счет «внутренних» валидационных фолдов.

Запуск тематической модели происходит 10 раз с разной инициализацией для каждого эксперимента.

5.2 Уравнивание весов модальностей

На рис. 1 демонстрируется улучшение качества решения вспомогательной задачи с помощью использования предложенной схемы уравнивания вкладов модальностей $\lambda_m = \mathbb{E}_{p(d)}(n_d^m)^{-1}$.

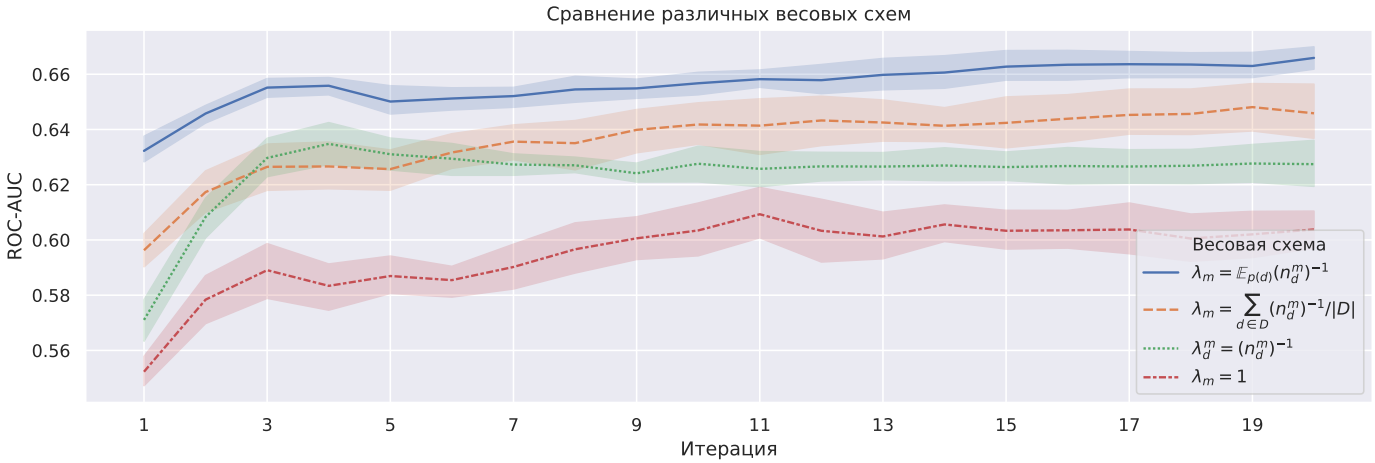

Рис. 1: Сравнение предложенной схемы уравнивания вкладов модальностей с единичными весами.

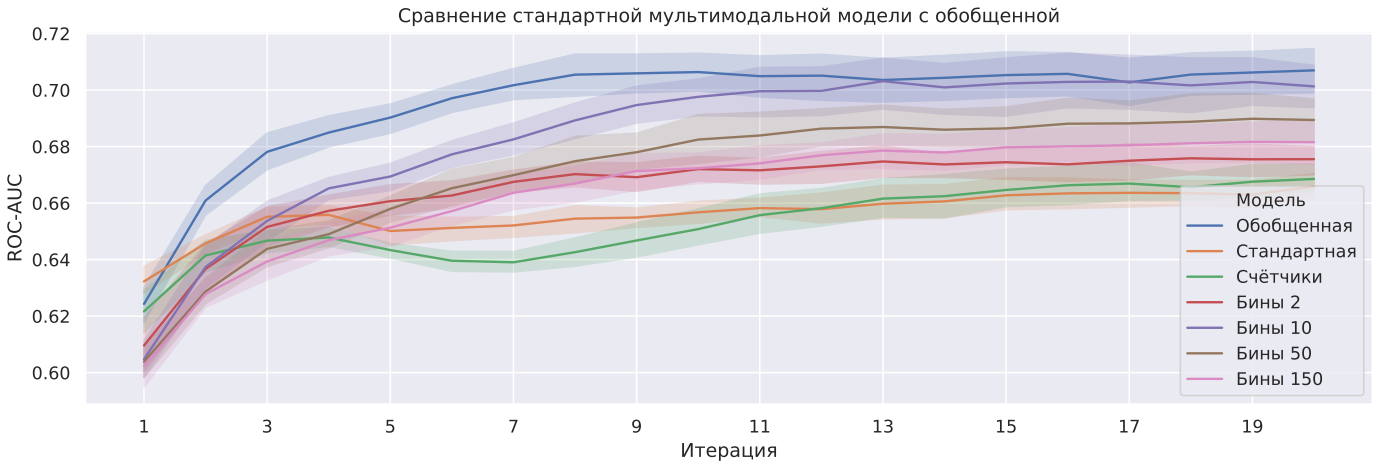

Рис. 2: Сравнение различных методик включения вещественной информации в тематическую модель.

5.3 Вещественные модальности

На рис. 2 демонстрируется зависимость качества решения вспомогательной задачи от итерации EM-алгоритма для следующих моделей:

- **обобщенная модель** — по предложенной в данной работе схеме к каждой теме добавляется своя гауссиана, параметры которой настраиваются в рамках EM-алгоритма
- **стандартная модель** — стандартная тематическая модель, не учитывающая вещественную информацию (суммы транзакций, в которых фирма выступает как продавец / покупатель)
- **счетчики** — для стандартных модальностей в качестве счетчиков используется логарифм суммы всех транзакций, в которых встречается рассматриваемый токен
- **бины** (корзины) — спектр значений вещественной модальности разбивается на n различных интервалов с помощью квантилей, затем значения заменяются номером соответствующего интервала, называемого *корзиной*

Использование сумм транзакций в качестве счетчиков n_{dw} не улучшает тематическую модель. Деление на корзины показывает результаты хуже, чем обобщенные модальности. Кроме того, для дискретизации на n корзин необходимо хранить $n \cdot |T|$ значений в матрице Φ, в то время как для обобщенной модели нужно только $2 \cdot |T|$ значений: математические ожидания и стандартные отклонения тематических гауссиан.

5.4 Влияние мультимодального моделирования на модальности

Мультимодальное тематическое представление позволяет сравнить раздельное обучение мономодальных тематических моделей с мономодальными тематическими представлениями Θ_m, полученными при обучении мультимодальной тематической модели.

Рис. 3 демонстрирует качество решения вспомогательной задачи для следующих тематических представлений:

- **мономодальные представления**
  - **мономодальное обучение:** для каждой модальности обучается своя тематическая модель
  - **мультимодальное обучение:** совместное обучение модальностей, мономодальные представления получены с помощью мультимодального разложения
- **мультимодальные представления** — с помощью мультимодального разложения можно получить выпуклую комбинацию мономодальных представлений, равную тематическим представлениям, полученным на M-шаге мультимодальной тематической модели
- **конкатенация** — мономодальные тематические представления конкатенируются в представление размерности $|M| \times |T|$, где $|M|$ — количество модальностей

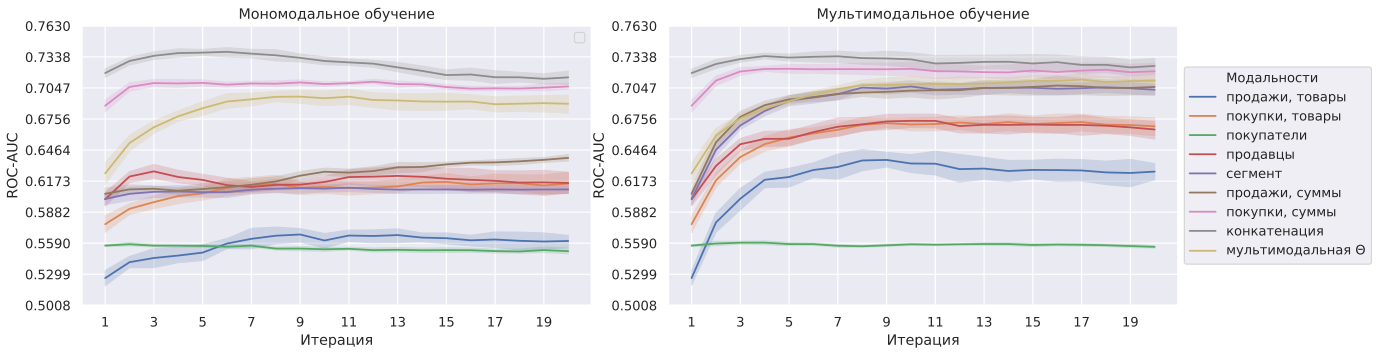

Рис. 3: Сравнение тематических представлений при мономодальном и мультимодальном обучении.

Данный эксперимент позволяет с помощью мультимодального разложения оценить разницу между раздельным и совместным обучением модальностей. Наблюдения:

1. Наилучшее качество показывает использование **конкатенации**, причем для мультимодальной модели конкатенация дает лучший результат.

2. Мультимодальные представления при совместном обучении показывают качество выше, чем мономодальные представления при раздельном, но хуже, чем лучшие мономодальные представления при совместном обучении.

5.5 Автоматический подбор весов модальностей

Мультимодальное разложение позволяет во время M-шага EM-алгоритма подбирать веса модальностей, оптимизирующих вспомогательный критерий. Важно отметить, что результирующее качество автоподбора определяется по отложенной выборке, не участвующей в подборе весов модальностей. При этом все эксперименты в данной работе демонстрируют качество модели на этой выборке.

Рестарты и инициализация. На рис. 4 сравниваются четыре различных варианта автоподбора:

- **Наличие рестарта**
  - Каждый шаг EM-алгоритма происходит инициализация весов модальностей
  - Инициализация весов модальностей происходит только на первой итерации, затем используются веса с прошлой итерации
- **Инициализация**
  - Инициализация весов модальностей происходит по логнормальному распределению, то есть логарифм весов модальностей имеет нормальное распределение
  - Веса инициализируются по предложенной в работе уравнивающей схеме:

$$\lambda_m = \mathbb{E}_{p(d)}(n_d^m)^{-1}$$

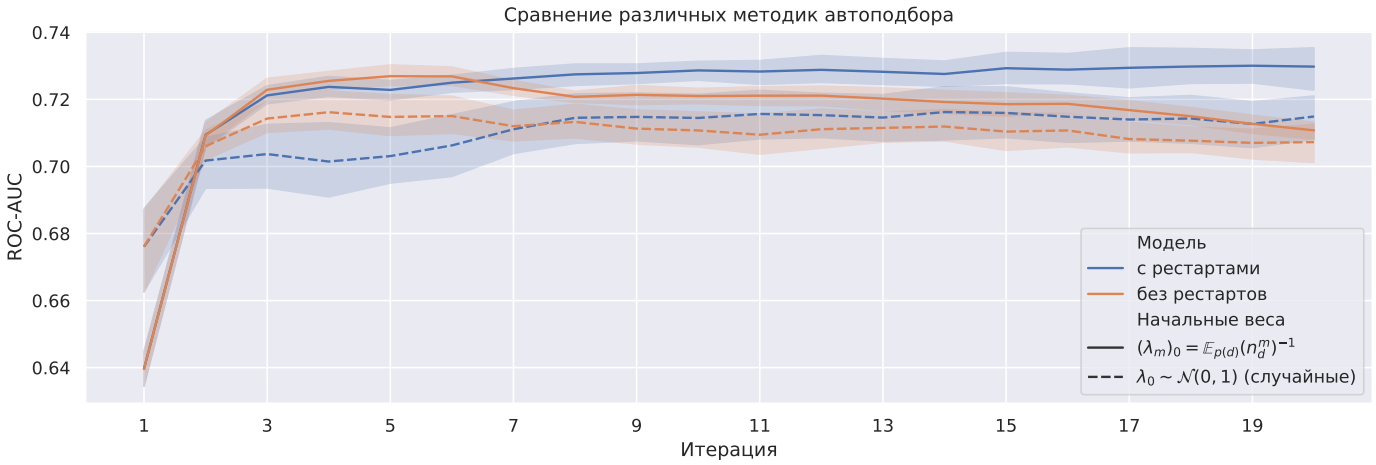

Рис. 4: Сравнение различных предложенных методик автоподбора весов модальностей.

Наилучший результат показывает модель с рестартами и инициализацией по уравнивающей схеме.

Веса модальностей при автоматическом подборе. На рис. 5 изображено изменение весов модальностей по мере обучения тематической модели с автоподбором весов.

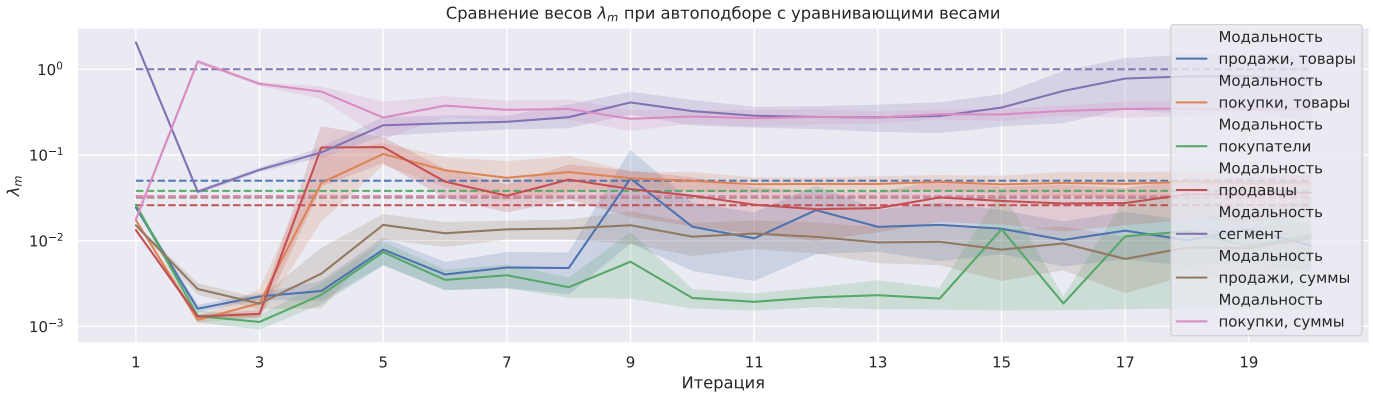

Рис. 5: Веса модальностей при автоподборе, логарифмическая шкала. Пунктирной линией обозначены соответствующие веса для уравнивающей схемы.

Прослеживается соответствие между качеством мономодальных представлений на правом графике рис. 3 и весами при автоподборе: чем выше качество мономодальных представлений, тем больше вес модальности при автоподборе.

Сравнение автоподбора с уравнивающей схемой. На рис. 6 демонстрируется улучшение качества решения задачи с помощью автоподбора.

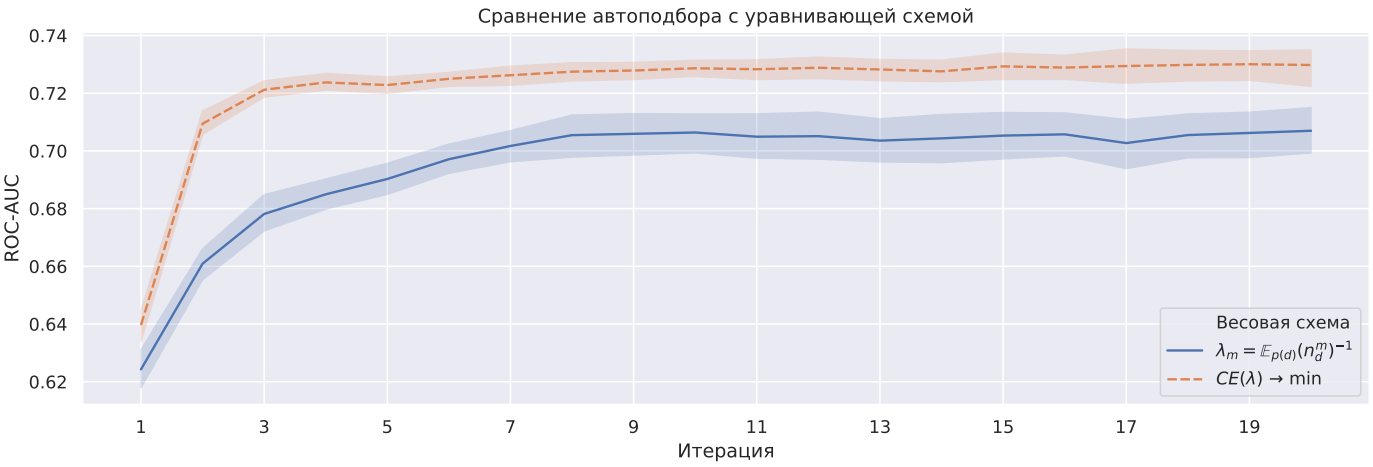

Рис. 6: Сравнение качества решения задачи при предложенной уравнивающей схеме и при автоподборе весов модальностей.

5.6 Итоговые результаты

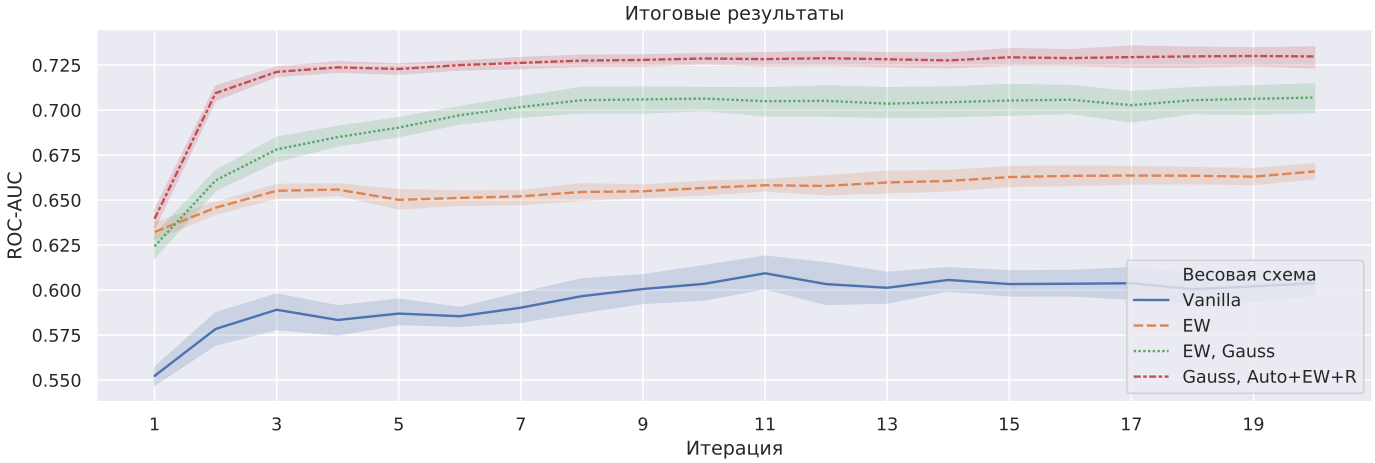

Рис. 7: Результаты экспериментов, все итерации.

В таблице 2 и на графике 7 приведены результаты экспериментов в виде сравнения «ванильной» тематической модели с инкрементальным добавлением предложенных модификаций:

- **Vanilla**: мультимодальная тематическая модель без вещественных модальностей и с единичными весами модальностей
- **EW**: $\lambda_m = \mathbb{E}_{p(d)}(n_d^m)^{-1}$
- **Gauss**: используются вещественные модальности
- **Auto+EW+R**: автоподбор весов модальностей с инициализацией по схеме EW на каждом шаге EM-алгоритма

Наилучший результат достигается в результате применения всех предложенных модификаций — использование вещественных модальностей с гауссианами и автоподбор весов модальностей путем решения вспомогательной задачи оптимизации, с повторной инициализацией предложенными уравновешивающими весами каждую итерацию EM-алгоритма.

Таблицы 3, 4, 5, 6 демонстрируют интерпретируемость полученной тематической модели.

Модель	Vanilla	EW	EW, Gauss	**Gauss, Auto+EW+R**
ROC-AUC	0.6153 ± 0.0108	0.6686 ± 0.0064	0.7126 ± 0.0109	$\mathbf{0.7356 \pm 0.0055}$

Таблица 2: Результаты экспериментов, лучшая итерация.

6 Заключение

Основные результаты данной работы:

- формализованы понятия вклада модальностей в мультимодальную тематическую модель, равного вклада модальностей и вклада модальностей, оптимального для вспомогательного критерия
- формализовано понятие вклада документа в мультимодальную тематическую модель, а также понятие равного вклада документов
- предложена и реализована методика уравнивания весов модальностей
- предложена и реализована методика автоматического подбора весов модальностей для вспомогательных критериев, которые можно сформулировать в виде классификации
- формализовано понятие вещественной модальности для тематической модели, использующее гауссианы
- разработана и реализована методика обучения вещественных модальностей с помощью EM-алгоритма
- проведены эксперименты, исследующие все предложенные модификации

Идеи для дальнейшего исследования:

- разработать и реализовать регуляризаторы для вещественных модальностей
- реализовать оптимизацию любых вспомогательных дифференцируемых критериев, принимающих на вход мономодальные тематические представления и веса модальностей
- разработать и реализовать регуляризаторы для весов модальностей при решении вспомогательных задач
- разработать и реализовать регуляризаторы для весов модальностей, основываясь на уже существующих регуляризаторах — декорреляция, разреженность и сглаживание
- разработать и реализовать механизм использования других распределений для вещественных модальностей

Результаты данной работы опубликованы в сборнике тезисов XXVII Международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов-2020».

Список литературы

[1] T. Hofmann. Probabilistic Latent Semantic Indexing // — Proceedings of the 22'nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — SIGIR '99. — New York, NY, USA : ACM, 1999. — C. 50–57.

[2] K. Vorontsov, O. Frei, M. Apishev et al. Bigartm: Open source library for regularized multimodal topic modeling of large collection // — International Conference on Analysis of Images, Social Networksand Texts — Springer. 2015. — C. 370–381

[3] A. P. Dempster, N. M. Laird, D. B. Rubin Maximum Likelihood from Incomplete Data via the EM Algorithm // — Journal of the Royal Statistical Society, Series B. 39.

[4] К. Хрыльченко. Тематическая модель экономической деятельности корпоративных клиентов банка // — Курсовая работа, 2019.

[5] Guolin Ke, Qi Meng, Thomas Finley et al. LightGBM: A Highly Efficient Gradient Boosting Decision // — Advances in Neural Information Processing Systems, 2017.

[6] J. H. Friedman. Greedy function approximation: a gradient boosting machine // — Annals of statistics, 2001. — C. 1189 – 1232.

[7] A. Paszke, S. Gross, F. Massa et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library // — Advances in Neural Information Processing Systems, 2019.

A Приложение

Продажи		**Покупки**	
Токен	**Вероятность**	**Токен**	**Вероятность**
весы	0.2329	сантехизделие	0.0497
банка	0.1568	хозтовары	0.0258
сантехнический	0.0776	гсм	0.0184
сантехник	0.0672	сантехнический	0.0159
всп	0.0274	труба	0.0135
смеситель	0.0246	пополнение	0.0133
битум	0.0207	бетон	0.0121
оборудование	0.0175	баланс	0.0116
изделие	0.0169	оплата	0.0112
сантехматериал	0.0162	арматура	0.0109
вса	0.0148	дорожный	0.0096
кран	0.0121	вр	0.0093
датчик	0.0096	весы	0.0092
фитинг	0.0084	щебень	0.0082
Кому продают		**У кого покупают**	
Токен	**Вероятность**	**Токен**	**Вероятность**
мир-весов	0.0308	вск-нева	0.0342
первый-весовой	0.0173	вираж-плюс	0.0220
ленвесторг	0.0167	деловые-линии	0.0174
максима	0.0151	ппр	0.0130
тк-вессервис	0.0144	мтс	0.0118
техмакс-трейд	0.0138	эго-инжиниринг	0.0090
форт-м4	0.0116	вираж	0.0083
газосфера	0.0093	бауцентр-рус	0.0077
современные-технолог...	0.0084	веста-регионы	0.0074
вск-нева	0.0077	битумойл	0.0060
инжиниринг	0.0071	рейс	0.0059
вессервис-нева	0.0069	адв-сервис	0.0055
глобвес	0.0069	мсервис	0.0054
сантехставопласт	0.0061	шинопторг	0.0053

Таблица 3: Описание темы t_1 — наиболее вероятные токены по распределению ϕ_{wt}.

Продажи		**Покупки**	
Токен	**Вероятность**	**Токен**	**Вероятность**
мониторинг	0.0836	часы	0.0278
опс	0.0801	автомобиль	0.0227
пож	0.0786	дать	0.0213
станция	0.0629	информационный	0.0191
обсл	0.0500	энергия	0.0190
абон	0.0421	карта	0.0137
пожарный	0.0377	кукуруза	0.0121
пожар	0.0273	мониторинг	0.0106
ремонтный	0.0251	добавка	0.0104
сист	0.0238	полистирол	0.0103
техн	0.0224	масло	0.0092
сигнал	0.0212	до	0.0091
пожарна	0.0198	стекло	0.0088
сигн	0.0149	ребёнок	0.0088
обслужить	0.0137	передача	0.0085

Кому продают		**У кого покупают**	
Токен	**Вероятность**	**Токен**	**Вероятность**
спейс	0.0222	ростелеком	0.0925
калинбар	0.0204	мегафон	0.0390
гранд-сервис	0.0142	вымпелком	0.0240
еврофудс-санкт-петер...	0.0140	севериконд	0.0120
бигбокс	0.0136	пищекомбинат-бежицки...	0.0101
ладога-спб	0.0116	костромская-сбытовая...	0.0091
оби-фц	0.0103	балтийский-лизинг	0.0087
алгос-фудс-2012	0.0099	красинформ	0.0073
фаворит-продукт	0.0098	красноярскэнергосбыт	0.0072
перспектива	0.0094	мобком	0.0068
алгос-фудс	0.0091	сергеевна	0.0061
галактика	0.0090	торговый-дом-астра	0.0057
риомаркет	0.0088	вест-колл-лтд	0.0054
гуд-фуд	0.0087	мобильные-телесистем...	0.0051
петрофуд	0.0086	сергиев-посад	0.0048

Таблица 4: Описание темы t_1 — наиболее вероятные токены по распределению ϕ_{wt}.

Продажи		Покупки	
Токен	**Вероятность**	**Токен**	**Вероятность**
мед	0.2278	мед	0.0851
медицинский	0.1565	питьевой	0.0167
осмотр	0.0598	медицинский	0.0156
медикамент	0.0555	транспортный	0.0151
медтовар	0.0369	товарный	0.0120
проведение	0.0266	закупка	0.0115
медосмотр	0.0215	объём	0.0108
предрейсовый	0.0205	шоколад	0.0107
водитель	0.0151	топливо	0.0105
препарат	0.0132	достижение	0.0103
тнп	0.0111	дизельный	0.0103
средство	0.0108	сервисный	0.0103
карта	0.0097	определённый	0.0100
сопутствующий	0.0094	лабораторный	0.0098
азс	0.0091	карта	0.0094

Кому продают		У кого покупают	
Токен	**Вероятность**	**Токен**	**Вероятность**
макс	0.0144	деловые-линии	0.0311
рива	0.0139	красноярскэнергосбыт	0.0221
фк-открытие	0.0122	вымпелком	0.0080
тамбовская-црб	0.0091	ростелеком	0.0076
восход	0.0074	юникредит-лизинг	0.0074
вск	0.0073	знак	0.0074
стегор	0.0070	прессол	0.0072
группа-ренессанс-стр...	0.0061	армадилло-бизнес-пос...	0.0066
офис-плюс	0.0060	карбогласс	0.0063
южный-двор-300	0.0059	петербургская-топлив...	0.0062
ресо-гарантия	0.0057	континент	0.0059
альянс-жизнь	0.0054	транспортная-лизинго...	0.0058
феникс	0.0051	заповедная-вода	0.0057
страховое-общество-г...	0.0045	престиж-интернет	0.0054
инженерный-консалтин...	0.0043	пауль-хартманн	0.0051

Таблица 5: Описание темы t_3 — наиболее вероятные токены по распределению ϕ_{wt}.

Продажи		**Покупки**	
Токен	**Вероятность**	**Токен**	**Вероятность**
автомобиль	0.1095	работник	0.0723
шиномонтаж	0.0721	автошина	0.0639
заявка	0.0708	лдсп	0.0415
автоуслуга	0.0388	одежда	0.0357
маршрут	0.0347	вино	0.0309
ту	0.0327	транспортный	0.0305
тэу	0.0304	водочный	0.0285
сжидить	0.0251	сжидить	0.0204
самореза	0.0232	комплектовать	0.0199
автотранспортный	0.0204	крепёж	0.0183
крепёж	0.0196	платный	0.0182
автомойка	0.0165	предоставление	0.0173
автотранспорт	0.0123	сотрудник	0.0145
мойка	0.0108	участок	0.0143
строительный	0.0107	энергия	0.0138
Кому продают		**У кого покупают**	
Токен	**Вероятность**	**Токен**	**Вероятность**
партс66	0.0545	кдв-групп	0.0320
сервер-авто+	0.0144	мишлен-русская-компа...	0.0182
сервер-опт	0.0137	кэш	0.0167
дельта	0.0116	автодор-платные-доро...	0.0134
злк	0.0098	иксора	0.0123
интерком	0.0096	феникс-плюс	0.0116
автопитер	0.0090	мегаком	0.0100
шате-м-плюс	0.0084	читаэнергосбыт	0.0098
крафтер	0.0080	пауэр-интернэшнл-шин...	0.0097
баянгол	0.0076	ростелеком	0.0084
тк-автопрайд	0.0073	красноярскэнергосбыт	0.0083
строй-лайн	0.0070	тк-мегаполис	0.0080
хай-кар	0.0064	амб-столица	0.0073
мокроусовский-коопзв...	0.0064	е100-онлайн	0.0068
строй-комплекс	0.0060	авто-траст	0.0066

Таблица 6: Описание темы t_4 — наиболее вероятные токены по распределению ϕ_{wt}.