© 2022 г. К.Я. ХРЫЛЬЧЕНКО (elightelol@gmail.com),
К.В. ВОРОНЦОВ, д-р физ.-мат. наук (vokov@forecsys.ru)
(ФИЦ «Информатика и управление» РАН, Москва)

ОПТИМИЗАЦИЯ ВЕСОВ МОДАЛЬНОСТЕЙ В ТЕМАТИЧЕСКИХ МОДЕЛЯХ ТРАНЗАКЦИОННЫХ ДАННЫХ[1]

Современные модели обработки естественного языка, такие как трансформеры, работают с мультимодальными данными. В данной работе исследуются мультимодальные данные с помощью мультимодального тематического моделирования над транзакционными данными корпоративных клиентов банка. Предлагается определение важности модальности для модели, на основе которого рассматриваются улучшения для двух сценариев моделирования: сохранение максимального количества информации с помощью балансирования модальностей и автоматический подбор весов модальностей для оптимизации вспомогательных критериев на основе тематических представлений документов.

Предлагается модель добавления численных данных в тематические модели в виде модальностей: каждой теме сопоставляется нормальное распределение с обучаемыми параметрами. Демонстрируются существенные улучшения по сравнению со стандартными тематическими моделями на задаче моделирования корпоративных клиентов банка. На основе тематических представлений клиентов банка прогнозируется 90-дневная просрочка по кредиту.

Ключевые слова: мультимодальное тематическое моделирование, транзакционные данные, классификация, прогноз просрочки по кредиту.

DOI: 10.31857/S0005231022120054, **EDN:** KRWGVF

1. Введение

Современные модели обработки естественного языка, такие как трансформеры [1, 2], принимают на вход мультимодальные данные: последовательность токенов, их позиций, а также сегментов в исходном тексте; для сессионных рекомендаций [3, 4] в качестве модальностей также используется большое количество дополнительной информации про активность пользователей. Модальность — это отдельный тип метаданных, присутствующий в моделируемых объектах, например, авторы или список литературы в статье. Несмотря на распространенность мультимодальных данных, методы для объединения мультимодальных данных в единое векторное представление объекта имеют

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 20-07-00936).

эвристический характер — складываются векторы, сформированные для отдельных модальностей. Это разумно только в предположении, что векторы модальностей находятся в едином семантическом векторном пространстве. Другим вариантом является конкатенация векторов с последующим применением многослойного персептрона.

Примером теоретически обоснованного объединения модальностей, имеющего вероятностную интерпретацию, является мультимодальное тематическое моделирование. В данной работе с помощью тематической модели составляются векторные представления корпоративных клиентов банка на основе мультимодальных транзакционных данных.

Основной гиперпараметр в мультимодальных тематических моделях — это веса модальностей. С их помощью задается влияние каждой модальности на итоговую тематическую модель. Плохо подобранные веса приводят к заглушению одних модальностей и доминации других; в худшем случае мультимодальная тематическая модель превращается в обычную тематическую модель с одной доминирующей модальностью. В разделе 3 приводятся объяснения, почему перебор весов модальностей плохо работает. В работе вводится определение важности модальности, на основе которого решаются следующие задачи: балансировка модальностей с целью сохранения наибольшего количества информации в модели и автоподбор весов модальностей для решения вспомогательной задачи, например, бинарной классификации для документов. В работе доказывается, что мультимодальные тематические представления документов можно представить в виде выпуклой комбинации тематических представлений, построенных для отдельных модальностей.

Предлагаемая в работе оценка оптимальных весов модальностей не требует перебора значений и повторных обучений тематической модели. Кроме того, демонстрируется существенное превосходство предлагаемых техник по сравнению с равными весами модальностей, являющимися, как правило, начальной точкой в любых экспериментах с тематическими моделями. Для демонстрации превосходства предложенных методов решается задача прогноза 90-дневной просрочки выплат по кредиту у корпоративных клиентов банка. В модели получаются интерпретируемые темы, которые можно сопоставить с различными видами экономической деятельности.

Численные характеристики, такие как количество цитирований статьи, обычно добавлются в модели с помощью различных эвристик. В нейронных сетях численные значения преобразуют в векторы с помощью многослойного персептрона, например, [4] преобразуют время совершения пользовательских действий в векторы с помощью линейного слоя. В данной работе предлагается интерпретируемый вероятностный метод для добавления численных данных в тематические модели. Каждой теме сопоставляется нормальное распределение с обучаемыми параметрами. Полученная адаптация EM-алгоритма для гауссовых модальностей в тематических моделях похожа на EM-алгоритм для смеси гауссиан [5].

Статья содержит следующую структуру. В разделе 2 дается математическая постановка задачи. Раздел 3 посвящен решению проблемы балансирования модальностей. В разделе 4 предлагается метод добавления вещественной информации в качестве модальностей в тематическую модель. Описание экспериментальной установки, а также результаты экспериментов содержатся в разделе 5. В заключительном разделе 6 подведены итоги работы и намечены дальнейшие исследования.

2. Тематическое моделирование

2.1. Обозначения

Тематическое моделирование позволяет сформировать интерпретируемые векторные представления для больших коллекций документов. Модель действует в предположении, что в данных присутствуют некоторые скрытые переменные, называемые темами.

Пусть $D = \{d_1, d_2, \dots\}$ — это коллекция сущностей, являющихся основными объектами моделирования, далее называемых документами. В свою очередь, документы состоят из токенов — в некотором смысле более «мелких», атомарных сущностей. Словарь W — это множество всех уникальных значений токенов.

Каждый документ $d \in D$ представляется как мультимножество $\{w_1, \dots, w_{n_d}\}$, где $w_i \in W \; \forall i$ и n_d — количество токенов в документе d.

Обозначим за $n := \sum\limits_{d \in D} n_d$ суммарную длину коллекции; n_{dw} — количество раз, которое токен w встретился в документе d.

2.2. Вероятностный латентно-семантический анализ

Основную гипотезу тематического моделирования, предложенную Томасом Хоффманом в модели PLSA [6], можно сформулировать следующим образом: каждой паре (токен w, документ d) в коллекции можно сопоставить некоторую **тему** t, при этом вероятность появления 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 некоторой вероятностной мерой P, где W, D, T — соответствующие множества токенов, документов и тем. Токены w и документы d являются наблюдаемыми переменными, а темы t — скрытыми.

Тогда условие «наличие токена w в документе зависит только от распределения документа по темам» представляется 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

$$p(w \mid d) = \sum_{t \in T} p(w \mid t, d) p(t \mid d) = \{1\} =$$

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

где $\Phi \in \mathbb{R}^{|W| \times |T|}$, $\Theta \in \mathbb{R}^{|T| \times |D|}$ — стохастические матрицы[2], являющиеся параметрами тематической модели.

Решение задачи максимизации логарифма правдоподобия наблюдаемой коллекции документов приводит к оптимальным значениям параметров Φ, Θ

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

Для решения используется EM-алгоритм [7]. Во время **E-шага** с помощью теоремы Байеса оценивается распределение скрытой переменной $p(t \mid d, w, \Phi, \Theta)$ для всех токенов $w \in W$ и документов $d \in D$:

$$p_{tdw} = p(t \mid d, w, \Phi, \Theta) = \big\{\text{теор. Байеса}\big\} =$$

$$= \frac{p(w \mid t, \Phi) p(t \mid d, \Theta)}{p(w \mid d, \Phi, \Theta)} = \frac{\phi_{wt} \theta_{td}}{\sum\limits_{s \in T} \phi_{ws} \theta_{sd}}. \tag{2}$$

На **M-шаге** решается задача $\mathbb{E}_{p_{tdw}} \log p(w, d, t', \mid \Phi, \Theta) \to \max\limits_{\Phi, \Theta}$

$$\begin{aligned} \phi_{wt} &= \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \right), \\ \theta_{td} &= \operatorname{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \right), \end{aligned} \tag{3}$$

где

$$\operatorname{norm}_{x \in X} f(x) := \frac{f(x)}{\sum\limits_{x \in X} f(x)}.$$

Решение можно получить с помощью ККТ [8].

На практике во время обучения модели не принято конструировать тензор p_{tdw} целиком в памяти. Достаточно итерироваться по документам и аккумулировать необходимые статистики. Такой алгоритм обучения тематических моделей называется рациональным EM-алгоритмом.

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

2.3. Мультимодальное тематическое моделирование

Рассмотрим ситуацию, в которой документ включает в себя данные разной природы, т.е. содержит разные модальности. Например, статья состоит из модальности слов и модальности авторов. Словарь для авторов состоит из всех возможных авторов, встречающихся в данной коллекции статей. Модальность m задается словарем W_m и распределением $p(w \mid t)$, $w \in W_m$, задаваемым стохастической матрицей $\Phi_m = (\phi_{wt} := p(w \mid t))$.

Чтобы подобрать оптимальные значения параметров, необходимо максимизировать взвешенную сумму логарифмов правдоподобия модальностей в коллекции

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

где $\lambda_m \geqslant 0$ — веса модальностей. Для упрощения записи обозначим за Φ конкатенацию матриц $\Phi_m \ \forall m$, в предположении, что словари модальностей не пересекаются: $W_i \cap W_j = \emptyset$, $i \neq j$.

E-шаг и обновление матрицы Φ выглядят аналогично PLSA (3). Обновление векторных представлений документов θ_{td} на M-шаге принимает вид

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

3. Взвешивание модальностей

При обучении мультимодальных тематических моделей распространена практика ручного перебора весов модальностей — изначально веса приравнивают к единичным $\lambda_m = 1 \, \forall m$, затем веса определенных модальностей итеративно увеличивают или уменьшают. У такого подхода две проблемы:

1. При изменении гиперпараметров необходимо переобучать модель. Более того, оценка качества полученной тематической модели является нетривиальной процедурой — ручная оценка интерпретируемости полученных тем или вычисление вспомогательного критерия на основе полученных векторов, зачастую требующее решения некоторой вспомогательной задачи. Простые процедуры, такие как вычисление перплексии, при решении практических задач не помогают найти оптимальные значения гиперпараметров.
2. Ручной подбор оптимальных весов модальностей — крайне не эффективная процедура. В разделе 3.1 доказывается, что оптимальные веса модальностей зависят от частотностей токенов модальностей в документах. Например, если в модели две модальности, одна из которых имеет в среднем сто токенов в документе, а другая один токен, необходимо увеличить вес модальности с одним токеном в сотню раз для примерно

равного влияния обеих модальностей на модель. Примеры таких частотностей очень распространены — текстовые модальности состоят из сотен токенов в документах, в то время как категориальные признаки кодируются одним токеном. Поэтому мультимодальные тематические модели могут быть нечувствительны к небольшим изменениям весов модальностей.

Так как проблема с долгим переобучением моделей несет исключительно вычислительный характер, стоит более детально обсудить проблему доминирующих и заглушаемых модальностей. Доминирующая модальность — это модальность, у которой в среднем много больше токенов в документе, чем у остальных модальностей. У заглушаемой модальности в среднем, наоборот, меньше токенов. При этом и заглушаемые, и доминирующие модальности могут содержать одинаково полезную информацию для тематической модели, а в некоторых ситуациях заглушаемая модальность может быть гораздо важнее. При наличии доминирующей модальности мультимодальная тематическая модель превращается в обычную тематическую модель с одной модальностью. Далее такие тематические модели с одной модальностью будут называться **унимодальными**.

Инициализация весов модальности сбалансированными значениями, предлагаемыми в данной работе, позволяет существенно улучшить тематическую модель. Под сбалансированными весами модальностей имеются в виду веса, при которых каждая модальность вносит равный вклад в векторные представления документов, изменения которых влияют на тематическую модель. В разделе 3.1 дается определение сбалансированных весов модальностей с помощью мультимодального разложения и предлагаются методики для их определения по статистикам исследуемой коллекции документов.

В отсутствие вспомогательной информации разумно предполагать, что вся имеющаяся информация одинаково важна, т.е. все модальности одинаково важны для модели вне зависимости от дальнейшего использования, будь то анализ полученных в коллекции тем или пользовательской активности. На практике часто доступна вспомогательная информация, с помощью которой можно оценить модель, такие как возраст пользователя, факт продолжительного использования сервиса пользователем, просрочка клиента банка по кредиту. При этом тематические векторные представления моделируемых документов можно использовать для решения вспомогательных задач, используя их в качестве входных данных.

Задачи обучения с учителем, решаемые на основе тематических векторных представлений, далее называем **целевыми задачами**. Тематическое моделирование с целевой задачей является наглядным примером признакоориентированного переноса обучения: модель предобучается на неразмеченных данных, затем используется как один из блоков для построения **целевой модели**, решающей целевую задачу. Разумно предположить, что важность разных типов входных данных тематической модели, т.е. модальностей, за-

висит от рассматриваемой целевой задачи. При предсказании возраста или половой принадлежности пользователя важна разная информация. Веса модальностей сильно влияют на итоговое качество решения целевой задачи.

Наивным подходом для автоподбора весов модальностей для целевой задачи является жадная пошаговая стратегия. Изначально выбирается одна модальность, затем происходит последовательное добавление новых модальностей. Предположим, что ищутся оптимальные веса модальностей $\lambda_1, \ldots, \lambda_{|M|}$ на симплексе $\Delta_{|M|} = \{\lambda_m \geqslant 0 \ \forall\, m \in M,\ \sum \lambda_m = 1\}$, оптимизируя вспомогательный критерий $J(\lambda)$. Пусть на шаге m зафиксированы веса первых m модальностей: $\lambda_1, \ldots, \lambda_m \in \Delta_m$. Тогда на шаге $m+1$ при добавлении модальности $m+1$ подбирается такое $\alpha \in [0,1]$, что значения

$$\lambda_1 := \alpha\lambda_1;\ \ldots;\ \lambda_m := \alpha\lambda_m;\ \lambda_{m+1} := (1-\alpha)$$

являются оптимальными весами модальности для критерия $J(\lambda(\alpha))$. Чтобы вычислить $J(\lambda(\alpha))$, необходимо обучить новую тематическую модель с весами модальностей $\lambda(\alpha)$. При наличии целевой задачи придется также переобучать и целевую модель, чтобы получить значение $J(\alpha)$. Для подбора α можно использовать неточный одномерный поиск, например, метод золотого сечения; применить методы оптимизации первого порядка, такие как градиентный спуск, нельзя из-за невозможности обратного распространения ошибки через ЕМ-алгоритм.

Большая часть целевых задач могут быть сформулированы как задачи оптимизации вспомогательного критерия $J(\Theta, w) \to \min$, где w — это параметры целевой модели. Предлагается методика для автоподбора весов модальностей для решения целевой задачи с дифференцируемым критерием во время М-шага, основанную на мультимодальном разложении (3.1). Данная методика не требует переобучения модели, неточного одномерного поиска или каких-либо других субоптимальных жадных пошаговых стратегий.

3.1. Мультимодальное разложение

Предположим, что сделан Е-шаг (2) ЕМ-алгоритма. Другими словами, зафиксируем условное распределение p_{tdw} скрытой переменной t. Пусть n_d^m — это количество токенов модальности m в документе d. Тогда

Теорема 1 (теорема о мультимодальном разложении). *В мультимодальной тематической модели на М-шаге векторные представления документа являются выпуклой комбинацией унимодальных*[3] *векторных представлений*

$$\theta_{td} = \sum_{m \in M} \tau_d^m \theta_{td}^m, \tag{4}$$

[3] Унимодальные представления получены с помощью М-шага для унимодальной тематической модели с одной модальностью, при том же значении p_{tdw}.

где

$$\tau_d^m = \frac{\lambda_m n_d^m}{\sum\limits_{m \in M} \lambda_m n_d^m}, \quad \theta_{td}^m = \mathrm{norm}_t \left(\sum_{w \in W_m} n_{dw} p_{tdw} \right).$$

Из данной теоремы, доказательство которой приводится в Приложении, следует несколько утверждений:

Следствие 1. Вектор $(\theta_{td}^m)_{t \in T}$ является унимодальным векторным представлением документа d и модальности m. Его можно получить с помощью M-шага (3) для стандартной тематической модели с одной модальностью. При $|M| = 1$ $\theta_{td} = \theta_{td}^m$. Мультимодальное векторное представление документа θ_{td} является комбинацией унимодальных векторных представлений θ_{td}^m.

Следствие 2. Коэффициент τ_d^m задает влияние модальности на мультимодальное представление документа d. При $\lambda_m = 1 \, \forall m$ у τ_d^m существует вероятностная интерпретация — это вероятность получить токен модальности m, если взять случайный токен из документа d.

Следствие 3. Из того, что $\sum\limits_{m \in M} \tau_d^m = 1$, $\lambda_m \geqslant 0$ следует, что τ_d лежит на симплексе. А значит, мультимодальное векторное представление документа θ_{td} является ***выпуклой*** *комбинацией унимодальных векторных представлений.*

Следствие 4. Веса модальностей $\lambda_m = (n_d^m)^{-1}$ уравнивают влияние различных модальностей на векторное представление документа d; получаются "равные вероятности" τ_d^m для всех модальностей $\lambda_m = \frac{1}{|M|} \, \forall m \in M$.

При использовании «небольшой» модальности вместе с «большой» модальностью, например, авторы и текст статьи, необходимо присваивать модальности автора в сотни раз больший вес, чем модальности текста статьи, чтобы модальность авторов оказывала видимое влияние на векторные представления документов.

Важности модальностей τ_d^m задаются отдельно для каждого документа $d \in D$ в коллекции. Чтобы сбалансировать модальности для каждого документа, необходимо вводить новую постановку задачи оптимизации, при которой каждому документу $d \in D$ соответствуют свои веса модальностей λ_d^m, $m \in M$, и присвоить этим весам модальностей значения $\lambda_d^m = (n_d^m)^{-1}$. Если оставаться в рамках стандартной постановки задачи, то предлагается следующая схема взвешивания модальностей:

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

Стоит отметить, что математическое ожидание по $p(d)$ увеличивает влияние длинных документов на итоговые веса модальностей.

Алгоритм 1 (EM-алгоритм с мультимодальным разложением).

1. **Вход:** Φ_k, Θ_k
2. вычислить p_{tdw} для $t \in T, d \in D, w \in W$, используя формулу 2
3. **для всех** $m \in M$ вычислить Φ^m и Θ^m, используя формулу 3
4. вычислить Θ_{k+1} как функцию от Θ^m, с помощью мультимодального разложения по формуле 4
5. вычислить Φ_{k+1} как конкатенацию $\Phi^1, \ldots, \Phi^m$
6. **Выход:** Φ_{k+1}, Θ_{k+1}

3.2. Целевые задачи

Сбалансированные веса модальностей позволяют избежать потери информации о менее частотных модальностях, нивелируя эффект их заглушения доминирующими модальностями, который в данной работе был явно выражен через мультимодальное разложение в теореме 1. Однако такие сбалансированные веса могут быть неоптимальными для целевых задач. Ниже предлагается метод для автоподбора весов модальностей, оптимальных для решения целевой задачи.

Алгоритм 1 позволяет обучить мультимодальную тематическую модальность с помощью мультимодальной декомпозиции. Он полностью эквивалентен стандартному мультимодальному рациональному EM-алгоритму с точки зрения получаемых параметров модели, однако он вычислительно менее эффективен: необходимо хранить векторные представления для всех модальностей $\Theta_1, \ldots, \Theta_m$, в то время как в рациональном алгоритме хранятся только финальные, мультимодальные векторные представления документов Θ. Тем не менее за счет разложения мультимодальных представлений в комбинацию унимодальных векторных представлений появляется возможность корректировать веса модальностей, не обучая новую тематическую модель. Финальные мультимодальные представления документов для заданного набора весов модальностей λ вычисляются с помощью важностей модальностей τ_d^m и векторов θ_{td}^m как выпуклая комбинация.

Пусть $J(\Theta)$ — вспомогательный оптимизационный критерий, который можно оптимизировать по λ, w:

$$\begin{cases} J(\Theta(\lambda), w) \to \min\limits_{\lambda, w}, \\ \lambda \succeq 0, \\ w \in D, \end{cases}$$

где

$$\Theta_{td} = \theta_{td}(\lambda) = \{4\} = \sum_{m=1}^{M} \frac{\lambda_m n_d^m}{\sum\limits_{m \in M} \lambda_m n_d^m} \theta_{td}^m,$$

$\lambda = (\lambda_1, \ldots, \lambda_M)$ — веса модальностей, D — допустимое множество значений w, например, $D = \{w : \langle w, 1_n \rangle = 1, w \succeq 0\}$.

Рассмотрим в качестве целевой задачи бинарную классификацию документов тематической модели. Каждый документ представляется как вероятностное распределение $(\theta_{td})_{t\in T}$. Будем оценивать вероятность положительного класса для документа d с помощью линейной модели

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

Предположим, что вероятность положительного класса зависит только от темы, тогда коэффициент w_t — это вероятность положительного класса для темы t:

$$p(y_d = 1 \,|\, d) = \sum_{t\in T} p(y_d = 1 \,|\, d, t)\, p(t \,|\, d) = \sum_{t\in T} p(y_d = 1 \,|\, t)\, p(t \,|\, d) = \sum_{t\in T} w_t \theta_{td}.$$

Используем мультимодальное разложение

$$\hat{y}_d = \sum_{t\in T} w_t \theta_{td} = \{4\} = \sum_{t\in T} w_t \sum_{m\in M} \frac{\lambda_m n_d^m}{\sum\limits_{m\in M} \lambda_m n_d^m} \theta_{td}^m =$$

$$= \sum_{m\in M} \frac{\lambda_m n_d^m}{\sum\limits_{m\in M} \lambda_m n_d^m} \hat{y}_d^m = \sum_{m\in M} \tau_d^m \hat{y}_d^m,$$

где $\hat{y}_d^m$ — это вероятность положительного класса для документа d при наличии в нем только модальности m.

Как было замечено в разделе 3, когда отсутствует информация о полезности тех или иных модальностей для решения целевой задачи, стоит считать их равноважными и использовать сбалансированные веса модальностей. Без балансировки модальности с большой средней мощностью[4] будут иметь большее влияние на тематическую модель. В случае классификации документов сбалансированные веса $\tau_d^m = 1/|M|$ приводят к усреднению оценок вероятности положительного класса m $\hat{y}_d^m$ по всем модальностям $m \in M$. Получается, что все модальности вносят равный вклад в предсказание; модальности с небольшой средней мощностью будут влиять на данные предсказания также, как и другие модальности, например авторы статьи и текст статьи.

В ситуации когда все-таки задан вспомогательный критерий J, подобрать оптимальные λ и w можно путем минимизации кросс-энтропии

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

[4] Мощность модальности — количество токенов этой модальности в документе.

Параметры λ, w можно подобрать для любого дифференцируемого критерия $J(\lambda, w)$ с помощью методов оптимизации первого порядка. В данной работе используется стохастический градиентный спуск. Преобразовать задачу в безусловную оптимизацию можно следующими заменами:

$$\lambda = \exp\hat{\lambda}, \quad w = \mathrm{softmax}(\hat{w}).$$

При оптимизации попеременно делаются шаги спуска по $\hat{\lambda}$ и $\hat{w}$.

4. Численные модальности

Предположим, что доступен некоторый численный признак или мультимножество численных характеристик для документа, например, суммы транзакций компании или возраст клиента. Предлагается моделировать эти численные значения с помощью нормального распределения.

Теорема 2 (теорема о гауссовых модальностях). *Пусть каждой теме $t \in T$ в тематической модели сопоставлена гауссиана $\mathcal{N}(\mu_t, \sigma_t^2)$ с обучаемыми параметрами. Тогда EM-алгоритм для обновления μ, σ^2 и всей тематической модели выглядят следующим образом:*

$$\mu_t = \sum_{d \in D} \sum_{x \in \mathbb{R}} \frac{n_{dx} p_{tdx}}{\sum\limits_{d \in D} \sum\limits_{x \in \mathbb{R}} n_{dx} p_{tdx}} x,$$

$$\sigma_t^2 = \sum_{d \in D} \sum_{x \in \mathbb{R}} \frac{n_{dx} p_{tdx}}{\sum\limits_{d \in D} \sum\limits_{x \in \mathbb{R}} n_{dx} p_{tdx}} (x - \mu_t)^2,$$

где n_{dx} — это количество раз, которое значение x встретилось в документе d, и p_{tdx} вычисляется во время E-шага: $p_{tdx} = p(t \mid d, x, \mu_t, \sigma_t)$. M-шаг выглядит идентично M-шагу в стандартной мультимодальной тематической модели.

Вывод EM-алгоритма для мультимодальной тематической модели с гауссовыми модальностями приводится в Приложении. Далее мультимодальные тематические модели с гауссовыми модальностями называются **обобщенными** мультимодальными тематическими моделями.

5. Эксперименты

5.1. Данные

Для экспериментов используются транзакционные данные корпоративных клиентов ПАО Росбанк. Документ порождается транзакциями клиента на заданном интервале времени. Товары и услуги, оказываемые в рамках транзакций, извлекаются из платежных поручений с помощью регулярных выражений. Активность компаний делится на продажи и покупки. Формируются семь различных модальностей: контрагенты (в двух ролях — продавцы и покупатели), товары и услуги (также в двух ролях), а также бизнес сегмент

клиента. В отличие от задач, основанных на текстовых коллекциях, возникают также численные модальности — суммы транзакций покупки и продажи.

Предложенные методы тестируются на задаче классификации документов: предсказывается 90-дневная просрочка выплат по кредиту на горизонте года у корпоративных клиентов банка. В качестве признакового пространства используются тематические векторные представления компаний, полученные из матрицы параметров тематической модели Θ. Количество тем в тематической модели зафиксировано: $|T| = 100$.

В качестве алгоритма классификации используется lightgbm [9] с зафиксированными параметрами: шаг обучения, равный значению 0.1, количество листьев, равное пяти, 50 шагов без улучшений до ранней остановки, и максимальное количество деревьев, равно 10 000. В качестве метрики классификации используется ROC-AUC.

Используется двойная кросс-валидация: данные делятся на 10 равных частей, называемых *внешними фолдами*. Каждый внешний фолд используется как отложенная выборка при обучении на остальных девяти фолдах, объединенных в одну обучающую выборку. Для ранней остановки градиентного бустинга полученная из девяти фолдов обучающая выборка делится на 10 *внутренних фолдов* и обучается 10 lightgbm моделей, каждая из которых использует свой внутренний фолд для ранней остановки. Стоит отметить, что внешние фолды никак не используются при обучении отдельных моделей или подбора гиперпараметров, поэтому полученные на них результаты можно считать валидными. Полученные на 10 внешних фолдах метрики качества усредняются.

Каждый эксперимент повторяется 10 раз с разными зафиксированными состояниями генератора случайности, результаты усредняются.

Кроме того, стоит еще раз отметить признако-ориентированный характер экспериментов — только для 83 тыс. документов в коллекции присутствует разметка, в то время как остальные данные являются неразмеченными и используются только для обучения тематической модели. При автоматическом подборе весов модальностей используется только разметка, попавшая в обучение итоговой целевой модели, т.е. градиентного бустинга.

5.2. Балансирование модальностей

Таблица 1 демонстрирует три различных схемы взвешивания модальностей:

1. $\lambda_d^m = (n_d^m)^{-1}$. Чтобы составить полностью сбалансированные веса модальностей для всех документов, необходимо изменить постановку задачи оптимизации и добавить веса λ_d^m для каждой модальности m и документа d и решить оптимизационную задачу

$$\sum_{\substack{m\in M,\\ d\in D}} \lambda_d^m \sum_{w\in W_m} n_{dw} \log \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

2. $\lambda_m = \dfrac{1}{|D|}\sum\limits_{d\in D}(n_d^m)^{-1}$. Данный подход предполагает усреднение обратных мощностей модальностей по всем документам в коллекции.
3. $\lambda_m = \mathbb{E}_{p(d)}(n_d^m)^{-1}$ корректирует прошлый подход с помощью вероятностей документов $p(d) = \dfrac{n_d}{\sum\limits_{d\in D} n_d}$.

5.3. Численные модальности

Разумно предположить, что суммы транзакций покупок, которые являются, по сути, потраченными компанией деньгами, важны для целевой задачи предсказания дефолта. Тем не менее результаты, представленные в табл. 4, показывают, что использование всей имеющейся информации улучшает метрики на целевой задаче по сравнению с лучшей унимодальной тематической моделью, использующей толькой суммы покупок.

Таблица 2 демонстрирует три метода добавления численной информации в модель:

1. Предложенные гауссовы модальности.
2. Счетчики. Во время M-шага (3), вместо использования частот n_{dw}, мы используем логарифмы сумм транзакций всех транзакций компании d, в которых встречался токен w. Идея заключается в том, что важность

Таблица 1. Сравнение различных схем взвешивания модальностей

Веса модальностей	ROC-AUC
$\lambda_m = 1$	$0{,}6153 \pm 0{,}0108$
$\lambda_d^m = (n_d^m)^{-1}$	$0{,}6422 \pm 0{,}0089$
$\lambda_m = \sum_{d\in D}(n_d^m)^{-1}/\lvert D\rvert$	$0{,}6546 \pm 0{,}0102$
$\lambda_m = \mathbb{E}_{p(d)}(n_d^m)^{-1}$	$0{,}6686 \pm 0{,}0064$

Таблица 2. Сравнение различных методов добавления численной информации. Стандартная модель не использует численную информацию. Все модели используют балансировку весов модальностей $\lambda_m = \mathbb{E}_{p(d)}(n_d^m)^{-1}$

Модель	ROC-AUC
стандартная	$0{,}6686 \pm 0{,}0064$
счетчики	$0{,}6716 \pm 0{,}0084$
150 корзин	$0{,}6864 \pm 0{,}0100$
50 корзин	$0{,}6936 \pm 0{,}0122$
100 корзин	$0{,}6945 \pm 0{,}0095$
10 корзин	$0{,}7082 \pm 0{,}0130$
гауссианы	$0{,}126 \pm 0{,}0109$

токенов для документа может быть оценена с помощью сумм транзакций, связанных с этим токеном.

3. Корзины. Все возможные значения численной характеристики разбиваются на n корзин, глядя на значения квантилей.

5.4. Решение целевой задачи

Во время М-шага, после вычисления векторных представлений θ_{td}^m для каждой модальности m, происходит корректировка весов модальностей. Исследуются два аспекта автоподбора:

1. Рестарты. Во время каждого М-шага мы переинициализируем веса модальностей. Отсутствие рестартов означает, что каждый М-шагов начинает корректировку весов модальностей с подобранных на прошлой итерации наилучших значений весов модальностей.
2. Инициализация. Перед автоподбором и, в случае рестартов, перед каждым М-шагом можно инициализировать веса модальностей сбалансированно $\lambda_m = \mathbb{E}_{p(d)}$, либо случайно, значениями из стандартного нормального распределения $\mathcal{N}(0, 1)$.

Веса модальностей λ и параметры линейного классификатора w попеременно оптимизируются шагами стохастического градиентого спуска. В качестве имплементации градиентного спуска используется библиотека для обучения нейросетей PyTorch [10], которая предоставляет возможность экспериментировать с более сложными классификаторами без необходимости собственноручно вычислять градиенты по параметрам.

Результаты представлены в табл. 3.

5.5. Результаты

Предлагаемые подходы сравниваются со стандартной мультимодальной тематической моделью, с лучшей унимодальной тематической моделью, а также с конкатенацией всех унимодальных моделей. Для каждой модальности обучается стандартная унимодальная тематическая модель. Лучшей унимодальной тематической моделью для предсказания дефолта является

Таблица 3. Сравнение различных методик автоподбора весов модальностей для решения задачи бинарной классификации на основе тематических векторных представлений

Рестарты	Сбалансированная инициализация	ROC-AUC
−	−	$0{,}7182 \pm 0{,}0083$
+	−	$0{,}7225 \pm 0{,}0111$
−	+	$0{,}7283 \pm 0{,}0044$
+	+	$0{,}7356 \pm 0{,}0055$

Таблица 4. Сравнение стандартной мультимодальной тематической модели, лучшей унимодальной модели, и конкатенации всех унимодальных представлений с предлагаемыми в работе методами

Модель	ROC-AUC
Стандартная	$0{,}6153 \pm 0{,}0108$
Сбалансированные веса	$0{,}6686 \pm 0{,}0064$
Гауссовы модальности	$0{,}7126 \pm 0{,}0109$
Лучшая унимодальная модель	$0{,}7131 \pm 0{,}0027$
Автоподбор весов модальностей	$0{,}7356 \pm 0{,}0055$
Унимодальная конкатенация	$0{,}7427 \pm 0{,}0057$

гауссова модальность с суммами транзакций, в которых компания-документ выступает как покупатель. Стоит отметить, что конкатенация унимодальных векторных представлений документов с точки зрения классификатора, особенно градиентного бустинга над деревьями, избавлена от проблем мультимодальной тематической модели. При конкатенации векторов сохраняется информация обо всех модальностях, отсутствует проблема доминирующих и заглушаемых модальностей. В то же время классификатор может использовать всю полезную информацию из модальностей, так как каждая модальность представлена набором признаков и у классификатора при выборе очередного признака для ветвления есть доступ к признакам всех модальностей.

Результаты, приведенные в табл. 4, приводят к следующим выводам:

1. Сбалансированные веса модальностей, численные модальности, автоподбор весов модальностей позволяют существенно повысить метрики на целевой задаче.
2. Лучшая мультимодальная тематическая модель, использующая предложенные техники, существенно превосходит лучшую унимодальную тематическую модель, что подчеркивает необходимость использования мультимодальных данных.
3. Сравнительно с конкатенацией унимодальных векторных представлений, предложенный подход сохраняет почти все качество, а значит и всю полезную информацию для решения целевой задачи. В некотором смысле мультимодальная тематическая модель — это алгоритм сжатия информации из матрицы совстречаемости токенов и документов для разных модальностей, а унимодальная конкатенация — это верхняя оценка на качество, которое можно сохранить при использовании мультимодальных данных.

Тем не менее мультимодальное представление, размерность которого в 7 раз меньше размерности конкатенации (100 против 700), представляет большую практическую ценность — современные масштабные системы для анализа транзакционных данных требуют хранения заранее посчитанных векторных представлений в быстрых key-value хранилищах, память в

которых является крайне ценным и ограниченным ресурсом. Кроме того, с точки зрения исследовательской ценности, мультимодальное представление имеет вероятностную интерпретацию и позволяет анализировать получившийся набор тем, а также оценивать схожесть различных документов и токенов с точки зрения различных векторных мер близости, в то время как конкатенация тематических представлений разных модальностей вероятностной интерпретации не поддается, а также требует дополнительных доработок для оценки близости различных элементов.

6. Заключение

В данной работе исследуется мультимодальность на примере тематических моделей, а также построение тематических моделей на основе транзакционных данных. Основной теоретический результат работы, мультимодальное разложение тематических представлений документов позволяет преобразовать мультимодальные векторные представления в выпуклую комбинацию унимодальных векторных представлений и оценить вклад модальностей в тематическую модель при заданных весах модальностей.

В данной работе предлагаются существенные улучшения для мультимодальных тематических моделей с помощью балансирования модальности и автоматического подбора весов модальностей для решения целевых задач. Вводится понятие численной модальности и демонстрируются существенные улучшения по сравнению с базовыми методами добавления численной информации в тематические модели. Все предложенные методы тестируются на задаче прогнозирования дефолта для корпоративных клиентов банка по транзакционным данным.

В качестве дальнейших исследований можно выделить два основных направления: регуляризацию весов модальностей, например, декорреляцию модальностей, а также расширение предложенных подходов для тематических моделей с регуляризацией [11].

ПРИЛОЖЕНИЕ

Доказательство теоремы 1.

Покажем, что вероятность θ_{td} темы t в документе d равна выпуклой комбинации θ_{td}^m:

$$\theta_{td} = \frac{\sum\limits_{m\in M}\lambda_m \sum\limits_{w\in W_m} n_{dw}p_{tdw}}{\sum\limits_{t\in T}\sum\limits_{m\in M}\lambda_m \sum\limits_{w\in W_m} n_{dw}p_{tdw}} = \frac{\sum\limits_{m\in M}\lambda_m \sum\limits_{w\in W_m} n_{dw}p_{tdw}}{\sum\limits_{m\in M}\lambda_m \sum\limits_{w\in W_m} n_{dw}\sum\limits_{t\in T} p_{tdw}} =$$

$$= \left\{\sum_{t\in T} p_{tdw} = 1\right\} = \frac{\sum\limits_{m\in M}\lambda_m \sum\limits_{w\in W_m} n_{dw}p_{tdw}}{\sum\limits_{m\in M}\lambda_m \sum\limits_{w\in W_m} n_{dw}} = \left\{\sum_{w\in W_m} n_{dw} = n_d^m\right\} =$$

$$= \frac{\sum\limits_{m\in M}\lambda_m \sum\limits_{w\in W_m} n_{dw}p_{tdw}}{\sum\limits_{m\in M}\lambda_m n_d^m} = \sum_{m\in M}\frac{\lambda_m n_d^m}{\sum\limits_{m\in M}\lambda_m n_d^m}\frac{\sum\limits_{w\in W_m} n_{dw}p_{tdw}}{n_d^m} = \sum_{m\in M}\tau_d^m\theta_{td}^m,$$

$$\text{где } \tau_d^m = \frac{\lambda_m n_d^m}{\sum\limits_{m\in M}\lambda_m n_d^m}, \quad \theta_{td}^m = \frac{\sum\limits_{w\in W_m} n_{dw}p_{tdw}}{n_d^m}.$$

Теорема 1 доказана.

Доказательство теоремы 2.

Разделим оптимизационный критерий для стандартных, $m \in M_s$, и гауссовых, $m \in M_g$, модальностей:

$$\mathcal{L}_s(\Phi,\Theta) := \sum_{m\in M_s}\lambda_m \sum_{t\in T}\sum_{d\in D}\sum_{w\in W_m} n_{dw}p_{tdw}\,\log\phi_{wt}\theta_{td},$$

$$\mathcal{L}_g(\mu,\sigma,\Theta) := \sum_{m\in M_g}\lambda_m \sum_{t\in T}\sum_{d\in D}\sum_{x\in\mathbb{R}} n_{dx}p_{mtdx}\log\mathcal{N}(x\,|\,\mu_{mt},\sigma_{mt}^2)\,\theta_{td}.$$

Тогда оптимизационная задача на M-шаге выглядит как:

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

и может быть поделена на три оптимизационных задачи.

1. Оптимизационная задача для Θ:

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

Решение задачи идентичной решению для стандартной мультимодальной тематической модели и выводится с помощью ККТ

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

2. Задача для Ф. Для модальности $m \in M_s$ и темы $t \in T$:

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

Решение задачи идентично решению задачи для стандартной мультимодальной тематической модели:

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

3. Нахождение оптимальных μ_t, σ_t^2. Для модальности $m \in M_g$ и темы $t \in T$

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

Достаточно приравнять производные к нулю и решить полученные уравнения

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

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

Теорема 2 доказана.

СПИСОК ЛИТЕРАТУРЫ

1. *Vaswani A., Shazeer N., Parmar N. et. al.* Attention is All you Need // Advances in Neural Information Processing Systems. 2017. V. 30.
2. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019. P. 4171–4186.
3. *Zhu W., Tao D., Cheng X. et. al.* BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer // CIKM. 2019. P. 1441–1450.
4. *Pavlovski M., Gligorijevic J., Stojkovic I. et. al* Time-Aware User Embeddings as a Service // Proceedings of SIGKDD Conference on Knowledge Discovery and Data Mining. 2020.

5. *Reynolds D.* Gaussian Mixture Models. Boston: Springer, 2009. P. 659–663.
6. *Hoffman T.* Probabilistic Latent Semantic Indexing // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99). 1999. P. 50–57.
7. *Dempster A.P., Laird N.M., Rubin D.B.* Maximum Likelihood from Incomplete Data via the EM Algorithm // J. Royal Statist. Soc. 1977. Series B 39. P. 1–38.
8. *Kuhn H.W., Tucker A.W.* Nonlinear programming // Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. 1950. P. 481–492.
9. *Ke G., Meng Q., Finley Th. et. al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural Information Processing Systems. 2017. V. 30.
10. *Paszke A., Gross S., Massa F. et. al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems. 2019. V. 30.
11. *Vorontsov V., Potapenko A.* Additive regularization of topic models // Mach. Learn. 101. 2015. P. 303–323.

Статья представлена к публикации членом редколлегии А.А. Лазаревым.

Поступила в редакцию 31.01.2022

После доработки 18.05.2022

Принята к публикации 29.06.2022