Автоматическая настройка параметров BigARTM под широкий класс задач.*

Гришанов А. В.[1]***, Булатов В. Г.***[1]***, Воронцов К. В.***[1]
grishanov.av@phystech.ru, bt.uytya@gmail.com, vokov@forecsys.ru
[1] Московский физико-технический институт

Открытая библиотека BigARTM позволяет строить тематические модели, используя широкий класс возможных регуляризаторов. При этом задача настройки коэффициентов оказывается весьма сложной. В данной статье мы ищем набор параметров, дающий «достаточно хорошие» результаты на широком классе задач, используя механизм относительных коэффициентов регуляризации и автоматический выбор N-грамм. Для эксперимента использовались наборы данных Victorian Era Authorship Attribution, 20 Newsgroups, МКБ-10. Модель с подобранными коэффициентами имеет качество не более чем на <X>% хуже «локально лучших моделей»

Ключевые слова: *тематическое моделирование, аддитивная регуляризация тематических моделей, PLSA, BigARTM.*

1 Введение

Тематические модели широко используются на практике для решения задач классификации и ранжирования документов, а также для разведочного поиска [3]. Одним из продвинутых инструментов в тематическом моделировании является библиотека BigARTM [4]. Она предоставляет широкий выбор для настройки модели, используя обширный класс регуляризаторов. Однако на практике популярнее остаются более простые методы, такие как LDA [1]. Во многом это связано со необходимостью аккуратного подбора параметров модели.

Предлагается проверить гипотезу о существовании конфигураций, хорошо работающих на широком классе задач.

Текущий подход к выбору параметров описан в работе [3]. Сначала подбирается один из регуляризаторов (например декоррелирования). Его значение находится приблизительно, затем в зависимости от цели исследуется следующий параметр или продолжает улучшаться текущий. Существенно то, что имеются различные регуляризаторы для различных итераций. Поэтому процесс перебора последовательный: добавляется один регуляризатор, оптимизируется, затем добавляется следующий. Из-за этого для оптимизации трудно использовать продвинутые методы, такие как байесовская оптимизация. Остаётся жадный или рандомизированный поиск.

При этом возможны 2 типа коэффициентов: абсолютные и относительные. Первые не переносятся между разными коллекциями, для них заранее трудно предположить даже порядок оптимальных значений. Вторые более универсальны, но требуют дополнительных вычислений и менее стабильны.

Цель данной работы — исследовать процесс подбора параметров и найти критерии, по которым можно выбирать начальные конфигурации BigARTM для широкого класса задач. Предлагается использовать относительные коэффициенты регуляризации и автоматический подбор N-граммм.

*Задачу поставил: Булатов В. Г. Консультант: Воронцов К. В.

2 Постановка задачи

Введем следующие обозначения: D — коллекция текстовых документов, состоящая из документов d; W словарь коллекции текстов, состоящий из термов w; T — множество тем, состоящее из тем t.

2.1 Будем считать, что выполнены следующие гипотезы.

Гипотеза о существовании тем. Каждое вхождение терма w в документ d связано с некоторой темой t из заданного конечного множества T . Коллекция документов представляет собой последовательность троек $\Omega_n = \{(w_i, d_i, t_i) : i = 1, \dots\ n\}$. Термы w_i и документы d_i являются наблюдаемыми переменными, темы t_i не известны и являются латентными (скрытыми) переменными.

Гипотеза «мешка слов». Порядок термов в документах не важен для выявления тематики, то есть тематику документа можно узнать даже после произвольной перестановки термов, хотя для человека такой текст потеряет смысл. Это предположение называют гипотезой «мешка слов» (bag of words). Порядок документов в коллекции также не имеет значения это предположение называют гипотезой «мешка документов».

Гипотеза условной независимости. Появление термов в документе d по теме t зависит от темы, но не зависит от документа d, и описывается общим для всех документов распределением $p(w|t)$:

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

2.2 Вероятностная тематическая модель порождения текста.

Согласно формуле полной вероятности и гипотезе условной независимости, распределение термов в документе $p(w|d)$ описывается вероятностной смесью распределений термов в темах $\varphi_{wt} = p(w|t)$ с весами $\theta_{td} = p(t|d)$

$$p(w|d) = \sum_{t\in T} p(w|t,d)p(t|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td}. \tag{2}$$

Вероятностная модель (2) описывает процесс порождения коллекции по известным распределениям $p(w|t)$ и $p(t|d)$. Этот процесс показан на рис. 1.

Рис. 1 Процесс порождения текстовой коллекции вероятностной тематической моделью (2): в каждой позиции i документа d_i сначала порождается тема $t_i \sim p(t|d_i)$, затем терм $w_i \sim p(w|t_i)$

2.3 Задача тематического моделирования — это обратная задача.

По заданной коллекции D требуется найти параметры φ_{wt} и θ_{td} , при которых тематическая модель (2) хорошо приближает частотные оценки условных вероятностей $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$. Распределение вида $p(t|x)$ будем называть тематикой объекта x. Можно говорить о тематике документа $p(t|d)$, терма $p(t|w)$, терма в документе $p(t|d,w)$. Целью тематического моделирования является определение тематики документов и связанных с ними объектов. Также требуется находить распределения $\varphi_{wt} = p(w|t)$, описывающие семантику каждой темы t словами естественного языка.

Равенство (2) можно переписать в матричном виде. В левой части равенства находится известная матрица частот термов в документах $F = (\hat{p}(w|d))_{W\times D}$. Ставится задача разложения матрицы F в произведение двух матриц Φ и Θ меньшего размера, таких, что

$$\Phi = (\varphi_{wt})_{W\times T},\ \varphi_{wt} = p(w|t) \text{ — матрица «термины-документы»}$$
$$\Theta = (\theta_{td})_{T\times D},\ \theta_{td} = p(t|d) \text{ — матрица «темы-документы»}$$

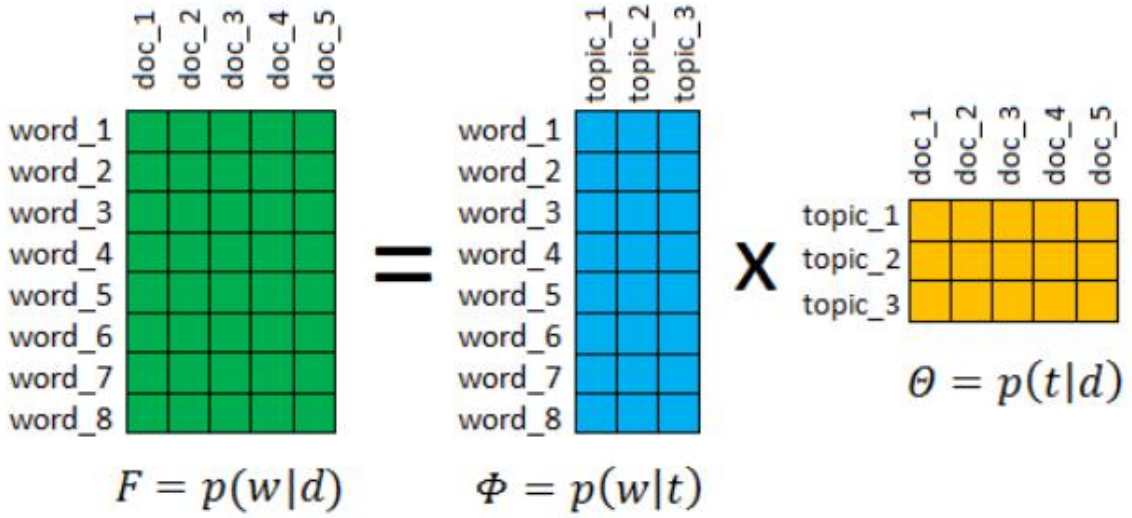

Рис. 2 Иллюстрация матричного разложения.

Поставленная задача ($F \approx \Phi\Theta$) эквивалентна поиску матриц Φ и Θ, максимизирующих следующий функционал правдоподобия:

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{3}$$

Заметим, что если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ — также является решением для всех невырожденных матриц S. Неоднозначность матричного разложения $F \approx \Theta\Phi$ порождает неоднозначность в выборе матриц из правой части равенства. Для решения данной проблемы, наложим на тематическую модель дополнительные требования. Модифицируем максимизирующий функционал:

$$L(\Phi,\Theta) + R(\Phi,\Theta) \to \max_{\Phi,\Theta} \tag{4}$$

$$R(\Phi, \Theta) = \sum_{j=1}^{n} \tau_j R_j(\Phi, \Theta) \tag{5}$$

где $R_j(\Phi, \Theta)$ — дополнительные требования к модели (регуляризаторы), τ_j — неотрицательные коэффициенты регуляризации. Полученная модель имеет название APTM (аддитивная регуляризация тематических моделей).

2.4 Коэффициенты регуляризации

TODO

2.5 Метрики качества

внутренние (intrinsic) и внешние (extrinsic)

TODO

2.6 Итоговая постановка

Рассмотрим набор датасетов $\{\mathfrak{D}_{ex}, \mathfrak{D}_{in}\}$, где $\mathfrak{D}_{ex} = \{\mathfrak{D}_j\}_{j=1}^{N_{ex}}$ имеют внешний критерий качества $S(\tau_j, \mathfrak{D}_j)$, а $\mathfrak{D}_{in} = \{\mathfrak{D}_j\}_{j=1}^{N_{in}}$ — только внутренние. Для каждого из первых найдём лучшие параметры.

$$\tau_{j_{best}} = \arg\min_{\tau} S(\tau, \mathfrak{D}_j), \;\; j = 1, \ldots, N_{ex} \tag{6}$$

Необходимо проверить гипотезу о том, что существуют коэффициенты регуляризации $\tau_{general}$, которые можно считать «универсальными», т.е. такие что выполнено:

$$\max_{j=1,\ldots,N_{ex}+N_{in}} \frac{S(\tau_{j_{best}}, \mathfrak{D}_j) - S(\tau_{general}, \mathfrak{D}_j)}{S(\tau_{j_{best}}, \mathfrak{D}_j)} \leqslant 5\% \tag{7}$$

Для поиска $\tau_{general}$ будем использовать $\mathfrak{D}_{ex}$, минимизируя следующий функционал:

$$\sum_{j=1}^{N_{ex}} \left(S(\tau_{j_{best}}, \mathfrak{D}_j) - S(\tau_{general}, \mathfrak{D}_j)\right)^2 \to \min \tag{8}$$

Гипотезу (7) будем проверять на $\{\mathfrak{D}_{ex}, \mathfrak{D}_{in}\}$. Для этого дополнительно введём следующие определения для наборов данных с внутренними критериями качества:

Определение 1. *Модель X* **не хуже** *чем модель Y, если по всем критериям X хуже не более чем на 5%.*

Определение 2. *Модель X* **лучше** *модели Y по k критериям, если по этим k критериям X лучше Y хотя бы на 5%.*

Цель работы — построить модель, которая **не хуже** чем PLSA [2] и **лучше** PLSA по нескольким критериям.

Литература

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 289–296, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[3] A.O. Ianina and K.V. Vorontsov. Multimodal topic modeling for exploratory search in collective blog. *Machine Learning and Data Analysis*, 2(2):173–186, 2016.

[4] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.