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 строит тематические модели, минимизируя регуляризованное правдоподобие с помощью EM-алгоритма. Подбор оптимальных параметров индивидуален для разных текстовых коллекций и требует работы специалиста. Для автоматизации этого процесса в данной статье мы ищем набор параметров, дающий достаточно хорошие результаты на широком классе задач. Предлагается перейти от относительного задания коэффициентов регуляризации к абсолютному. 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 два 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.

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

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

2 Постановка задачи

Заданы: набор текстовых коллекций $\mathfrak{D} = \{D_i\}_{i=1}^N$, где каждая коллекция D состоит из документов d, словари коллекциий текстов W, состоящих из термов w, и множества тем T, состоящих из тем t.

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.

Пусть задан критерий качества тематического моделирования S. Цель данной работы состоит в решении следующей задачи:

$$S \to \max_{\mathfrak{D}} \tag{1}$$

3 Задача тематического моделирования

Согласно формуле полной вероятности и гипотезе условной независимости:

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

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

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

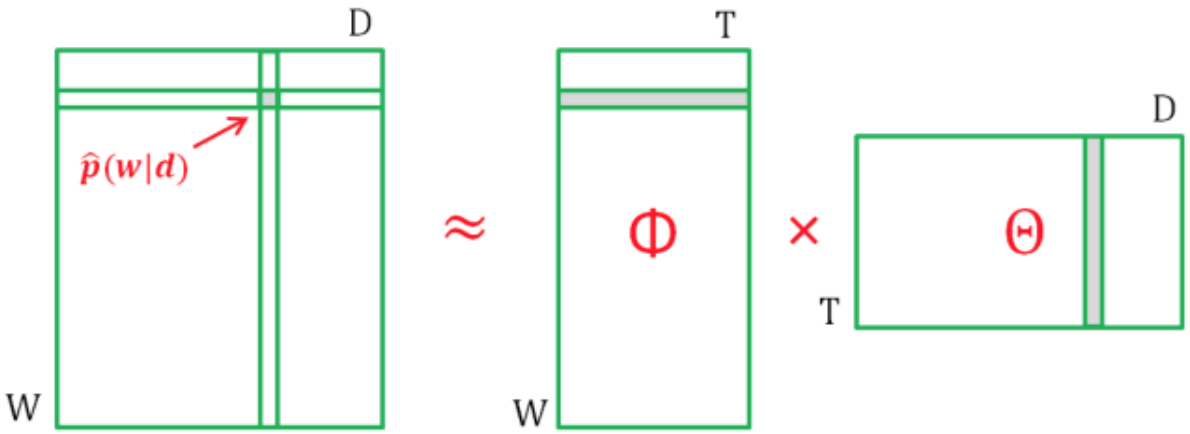

Рис. 1 Иллюстрация матричного разложения.

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

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

Решение данной задачи неединственно, что 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. Полученный подход к построению тематических моделей имеет название АРТМ (аддитивная регуляризация тематических моделей).

4 Итоговая постановка

Для оценивания тематической модели используют два типа критериев качества

Определение 1. *Внутренние критерии оценивают качество построенной модели по итоговым матрицам Φ и Θ модели.*

Определение 2. *Внешние критерии измеряют качество полученных предсказаний.*

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

Определение 4. *Модель X* **лучше** *модели Y по k критериям, если по этим k критериям X лучше Y хотя бы на 5%.*

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

5 Относительные коэффициенты регуляризации

Формула М-шага, сглаживающего ($\tau > 0$) или разреживающего ($\tau < 0$) распределение φ_{wt}:

$$\varphi_{wt} = \underset{w \in W}{\text{norm}} (n_{wt} + \tau) \tag{9}$$

Интуитивный смысл этого преобразования прост: мы либо «притягиваем» Фи к равномерному распределению $\beta = \frac{1}{|W|}$, либо "отталкиваем" её от него же (возможно, даже зануляя при этом какие-то компоненты). Оказывается, можно провести репараметризацию, которая строго это продемонстрирует.

Пусть $\beta = \frac{1}{|W|}$ — равномерное распределение, а текущие значения n_{wt} и $\tau \in \mathbb{R}$ таковы, что на этой итерации М-шага зануления компонент не происходит (то есть либо $\tau > 0$, либо $\tau < 0$, но $n_{wt} + \tau > 0$). Тогда операцию положительной обрезки можно проигнорировать:

$$\varphi_{wt} = \underset{w \in W}{\text{norm}} (n_{wt} + \tau) = \frac{n_{wt} + \tau}{\sum\limits_{w \in W} n_{wt} + \tau} = \frac{n_{wt} + \tau}{n_t + \tau |W|} \tag{10}$$

представим это выражение, как выпуклую комбинацию распределений $\frac{n_{wt}}{n_t}$ (оценки максимума правдоподобия) и $\frac{1}{|W|}$ (равномерного распределения)

$$\frac{n_{wt} + \tau}{n_t + \tau |W|} = (1 - \lambda) \frac{n_{wt}}{n_t} + \lambda \frac{1}{|W|} \Rightarrow \tau = \frac{n_t \lambda}{(1 - \lambda)|W|} \tag{11}$$

Значит, сглаживание Фи можно трактовать, как нахождение компромисса между $\varphi_{wt} = \frac{n_{wt}}{n_t}$ и $\varphi_{wt} = \frac{1}{|W|}$.

Допустим, мы хотим провести регуляризацию так, чтобы φ_{wt} на 50% состояла из оценки максимума правдоподобия, и на 50% из априорного распределения $\frac{1}{|W|}$. Для этого достаточно вычислить τ по формуле и подставить в модель.

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.