Автоматическая настройка параметров APTM под широкий класс задач

Е. В. Иванова[1], С. Н. Матвеева[2], Т. А. Голубева[3], А. В. Трусов[4], В. В. Черноног[5], М. В. Царицын[6]

Аннотация: В работе рассматривается задача тематического моделирования. Тематическое моделирование нашло своё применение в таких областях как машинное обучение и обработка естественного языка. В работе используется широко известная библиотека BigARTM, использование которой требует настройки большого числа параметров. В работе рассматривается возможность нахождения универсального набора значений параметров для широкого класса задач. Для нахождения этих значений, используется метод <...>. Для оценки качества используется критерий <...>. Полученная, с помощью подобранных коэффициентов, ошибка не больше чем на <X>% больше чем в "локально лучших моделях".

Ключевые слова: тематической моделирование, ARTM, BigARTM, глобальная оптимизация.

1 Введение

Тематическое моделирование — способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов[4]. Тематическая модель (англ. topic model) коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова образуют каждую тему. Тематические модели широко используются на практике для решения задач классификации и ранжирования документов, а также для разведочного поиска[6]. Открытая библиотека bigARTM[3] позволяет строить тематические модели, используя широкий класс возможных регуляризаторов[5]. Однако такая гибкость приводит к тому, что задача настройки коэффициентов оказывается очень сложной.

[1]Московский физико-технический институт, ivanova.ev@phystech.edu
[2]Московский физико-технический институт, matveeva.sn@phystech.edu
[3]Московский физико-технический институт, golubeva.ta@phystech.edu,
[4]Московский физико-технический институт, trusov.av@phystech.edu,
[5]Московский физико-технический институт, chernonog.vv@phystech.edu,
[6]Московский физико-технический институт, tsaritsyn.mv@phystech.edu,

Эту настройку можно значительно упростить, используя механизм относительных коэффициентов регуляризации и автоматический выбор N-грамм. В работе проверяется гипотеза о том, что существует универсальный набор относительных коэффициентов регуляризации, дающий "достаточно хорошие"результаты на широком классе задач. Дано несколько датасетов с внешним критерием качества (классификация документов по категориям и ранжирование). Находятся лучшие параметры для конкретного датасета, дающие "локально лучшую модель". Ищется алгоритм инициализации bigARTM, производящий тематические модели с качеством, сравнимым с "локально лучшей моделью"на её датасете. Критерий сравнимости по качеству: на данном датасете качество "универсальной модели"не более чем на 5% хуже, чем у "локально лучшей модели".Производится сравнение с другими более простыми моделями: PLSA(Probabilistic latent semantic analysis)[2] и LDA(Latent Dirichlet allocation)[1]

2 Постановка задачи

2.1 Рассматриваемые датасеты

2.1.1 Victorian Era Authorship Attribution Data Set

Рассматривается датасет Victorian Era Authorship Attribution Data Set. В данном датасете рассматриваются авторы, удовлетворяющие следующим критериям:

- англоязычные авторы
- авторы, у которых не менее 5 книг
- авторы XIX века

Всего рассматривается около 50 авторов. В выбранных текстах были отброшены первые и последние 500 слов, чтобы избавиться от несущественной информации, такой как имя и другая информация об авторе, название книги и тд. Все тексты были изучены для получения общего числа уникальных слов и их частот. После этого из тектов взяли наиболее часто всречаемые 10000 слов, которые были пронумерованы в порядке убывания частот.

2.1.2 Twenty Newsgroups Data Set

Датасет представляет собой набор из примерно 20 000 документов, разделенных почти равномерно по 20 различным группам новостей. Статьи являются типичными примерами публикаций и, таким образом, имеют заголовки, включая темы, подписи и цитируемые части других статей.

2.1.3 МКБ-10

МКБ-10 — Международная классификация болезней 10-го пересмотра. На январь 2007 года является общепринятой классификацией для кодирования медицинских диагнозов, разработана Всемирной организацией здравоохранения. МКБ-10 состоит из 21 класса (раздела), каждый из которых содержит рубрики с кодами заболеваний и состояний.

2.1.4 2 habr classification

Необходимо по тексту определить: лучше публиковать его в блоге на Хабрахабр или на Geektimes, другими словами нужно научить алгоритм отличать статьи одного блога от другого. Подразумевается, что текст технический и релевантен тематике данных блогов. В качестве исходных данных используются два json файла с 1000 текстами с каждого из этих двух сайтов.

2.2 Метрика

В качестве меры важности слов в контексте документа используется tf-idf.

TF — это частотность термина, которая измеряет, насколько часто термин встречается в документе. В длинных документах термин может встретиться в больших количествах. Поэтому применяют относительные частоты — делят количество раз, когда нужный термин встретился в тексте, на общее количество слов в данном тексте.

IDF — это обратная документная частота. Считается как логарифм от общего количества документов, делённого на количество документов, в которых встречается термин.

2.3 Обзор тематических моделей

2.3.1 Постановка задачи тематического моделирования

Рассмотрим словарь терминов W из элементов которого складываются документы, и коллекцию D документов $d \subset W$. Для каждого документа d известна его длина n_d и количество n_{dw} использований каждого термина w. Необходимо найти параметры вероятностной порождающей тематической модели, то есть представить вероятность появления p(w|d) слов в документе в виде:

$$p(w|d) = \sum_{t \subset T} \phi_{wt} \Theta_{td}, \tag{1}$$

где ϕ_{wt}= p(w|t) — вероятности терминов w в каждой теме t, Θ_{td} = p(t|d) — вероятности тем t в каждом документе d.

Порождающая модель описывает процесс построения коллекции по ϕ_{wt} и Θ_{td}. Тематическое моделирование представляет собой обратную задачу: по наблюдаемой коллекции необходимо понять, какими распределениями ϕ_{wt} и Θ_{td} она могла бы быть получена.

2.3.2 LDA

При обработке на естественном языке скрытое распределение Дирихле (LDA) представляет собой генеративную статистическую модель, которая позволяет объяснять группы наблюдений ненаблюдаемыми группами, которые объясняют, почему некоторые части данных схожи. Например, если наблюдения представляют собой слова, собранные в документы, в нем говорится, что каждый документ представляет собой смесь небольшого количества тем и что присутствие каждого слова связано с одной из тем документа. LDA - пример тематической модели.

В LDA каждый документ может рассматриваться как смесь различных тем, где каждый документ считается набором тем, которые ему назначаются через LDA. Это идентично вероятностному латентному семантическому анализу (pLSA), за исключением того, что в LDA предполагается, что в распределении темы имеется редкий Dirichlet. Редкие Dirichlet priors кодируют интуицию, что документы охватывают только небольшой набор тем и что темы часто используют только небольшой набор слов. На практике это приводит к лучшему рассогласованию слов и более четкому присваиванию документов тем. LDA является обобщением модели pLSA, которая эквивалентна LDA при равномерном распределении Дирихле.

2.3.3 PLSA

Вероятностный латентный семантический анализ (PLSA), также известный как вероятностное латентное семантическое индексирование (PLSI, особенно в кругах поиска информации), является статистическим методом анализа двухмодовых и совпадающих данных. Фактически, можно получить низкоразмерное представление наблюдаемых переменных в терминах их близости к некоторым скрытым переменным, как и в латентном семантическом анализе, из которого развилась PLSA.

По сравнению со стандартным латентным семантическим анализом, который проистекает из линейной алгебры и сокращает таблицы возникновения (обычно через разложение сингулярных значений), вероятностный латентный семантический анализ основан на разложении смеси, полученном из модели скрытого класса. В этом случае регуляризатор не используется.

Список литературы

[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*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[3] 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.

[4] Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.

[5] КВ Воронцов. Вероятностное тематическое моделирование. *Москва*, 2013.

[6] Анастасия Олеговна Янина and Константин Вячеславович Воронцов. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. *Машинное обучение и анализ данных*, 2(2):173–186, 2016.