

Метаобучение тематических моделей классификации

А. С. Ватолин¹, Ю. А. Сердюк², К. В. Воронцов¹

vatolinalalex@gmail.com; masyes@mail.com; vokov@forecsys.ru

¹ Московский физико-технический институт, Москва, Россия

Одним из возможных применений вероятностной тематической модели является построение модели не только для текста, но и для имеющихся метаданных (модальностей). Это позволяет более точно определять темы документов, а также предсказывать пропущенные метаданные по имеющимся. Каждая модальность имеет свой вес, который задается вручную и отражает меру влияния данной модальности на темы документов. В данной работе исследуются эвристики для начальной инициализации весов модальностей. Получение эвристики позволит полностью отказаться от перебора весов модальностей по сетке или же уменьшить количество вариантов перебора. Для оценки весов используется мера взаимной информации.

Ключевые слова: *вероятностное тематическое моделирование; Мультимодальное тематическое моделирование; BigARTM*

1 Введение

Вероятностное тематическое моделирование - способ построения модели текстовых документов, которая определяет к какой теме относится каждый документ и какие слова образуют тему. Тематическое моделирование применяется в информационном поиске [4], для классификации [3] и суммаризации текстов [7], а также для ранжирования статей [1]. Одним из продвинутых инструментов, который реализует все из перечисленных выше инструментов, является библиотека BigARTM [6]. Она обладает обширным набором параметров для настройки модели, а также различными регуляризаторами.

В данной статье внимание сконцентрировано на задаче классификации, а именно предлагаются эвристики для оптимальной инициализации весов модальностей в тематической модели. Вводится предположение о том, по статистикам исходной коллекции текстов можно вывести оценку. Текущие подходы к выбору весов модальности сводятся к двум методам: перебор параметров по сетке или задание константного имперического значения. В работе [2] определение весов осуществляется перебором параметров по сетке методом проб и ошибок. Выбор лучшего набора параметров осуществлялся по критериям перплексии, разреженности и качества тематического поиска. Также в работе [8] предлагается ввести следующее правило: дополнительные модальности не должны увеличивать перплексию основной модальности. Введение этого правила не избавляет от перебора по сетке, а лишь изменяет правила выбора лучшего значения для весов. В статье [5] веса модальностей задаются вручную из соображений важности каждой из модальностей для модели.

Одной из близких к данной задаче является настройка весов регуляризаторов вероятностной тематической модели [6]. Так как задача тематического моделирования является многокритериальной, то одновременно при обучении модели может использоваться несколько регуляризаторов, сбалансированных с помощью весов. В ходе экспериментов было установлено, что весов регуляризаторов зависят от параметров выборки, таких как размер коллекции, мощность словаря, средняя длина документов. Для избавления от необходимости перенастройки весов при изменении параметров коллекции вводятся относительные коэффициенты регуляризации.

Подбор параметров осуществляется на выборках R8, R52, AG's news, Ohsumed, 20NG, DBPedia, IMDb, Amazon 2, Yelp 5, Sogou News.

2 Постановка задачи

Данный документ демонстрирует оформление статьи, подаваемой в электронную систему подачи статей <http://jmla.org/papers> для публикации в журнале «Машинное обучение и анализ данных». Более подробные инструкции по стилевому файлу `jmla.sty` и использованию издательской системы \LaTeX 2 ϵ находятся в документе `authors-guide.pdf`. Работу над статьёй удобно начинать с правки \TeX -файла данного документа.

Обращаем внимание, что данный документ должен быть сохранен в кодировке UTF-8 without BOM. Для смены кодировки рекомендуется пользоваться текстовыми редакторами Sublime Text или Notepad++.

2.1 Название параграфа

Разделы и параграфы, за исключением списков литературы, нумеруются.

3 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

Литература

- [1] Dongsheng Duan, Yuhua Li, Ruixuan Li, Rui Zhang, and Aiming Wen. Ranktopic: Ranking based topic modeling. In *2012 IEEE 12th International Conference on Data Mining*, pages 211–220. IEEE, 2012.
- [2] AO Ianina and KV Vorontsov. Multimodal topic modeling for exploratory search in collective blog. In *Proc. 11th Int. Conf. on Intelligent Data Processing: Theory and Applications*, pages 186–187, 2016.
- [3] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88(1-2):157–208, 2012.
- [4] Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456, 2011.
- [5] Воронцов К. В. Ефимова И. В. Иерархическая мультимодальная тематическая модель коллекции научно-популярных текстов.
- [6] Воронцов К.В. Вероятностное тематическое моделирование: теория, модели и проект bigartm. 2020.
- [7] Бакиева А. М/. Создание системы автоматического реферирования научных текстов. *Вестник Новосибирского государственного университета. Серия: Информационные технологии*, 16(3), 2018.
- [8] Матвеев И. А. Никитин Ф. А., Воронцов К. В. Применение мультимодальных тематических моделей к анализу транзакционных данных банков.

Поступила в редакцию