Метаобучение тематических моделей классификации

А. С. Ватолин[1], *Ю. А. Сердюк*[2], *К. В. Воронцов*[1]
vatolinalex@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 резуляризаторами.

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 парамеров осуществлялся по критериям перплекции, разреженности и качества тематического поиска. Также в работе [8] предлагается ввести следующее правило: дополнительные модальности не должны увеличивать перпрелксию 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 парамеров 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

Пусть задано $\mathfrak{D} = \{D_i\}_{i=1}^N$ - множество (коллекция) текстовых документов. Каждая коллекция D состоит из документов d. T - множество тем, M - множество модальностей. $W^1, ..., W^m$ - словари термов, соответствующие модальностям $m \in M$. Термами могут быть слова, нормальные формы слов, словосочетания и т.д. в зависимости от предобработки исходных данных. Каждый документ $D \in \mathfrak{D}$ представлен в виде последовательности n_d термов $w_1, ..., w_{n_d} \in W^m \ \forall m \in M$, где каждому терму ставится в соответствие число его вхождений n_{dw}.

При построении мультимодальной тематической модели используются следующие обозначения: $p(w|t)$ - вероятность появления терма w в теме t, $p(t|d)$ - вероятность появления темы t в документе d.

Тогда вероятность появления терма произвольной модальности в документе может быть выражена следующим образом:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}, \ w \in W_m, d \in D \tag{1}$$

Для каждой модальности m ставится в соответствие матрица $\Phi_m = (\varphi_{wt})_{W_m \times T}$. При объединении матриц каждой модальности Φ_m в одну матрицу образуется $W \times T$-матрица Φ.

Мультимодальная модель строится путем максимизации взвешенной суммы логарифмов правдоподобия модальностей и регуляризаторов.

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta} \tag{2}$$

Предсказание делается следующим образом:

$$p(c|d) = \sum_{t \in T} \varphi_{ct}\theta_{td} \tag{3}$$

Документ d относится к классу c, если $p(c|d) >= \gamma_c$

Качество классификации измеряется с помощью метрики $f1$ с микро усреднением

$$f1_{macro} = \frac{2}{C} \sum_{c \in C} \cdot \frac{precision_c \cdot recall_c}{precision_c + recall_c} \tag{4}$$

Для определения оптимальных весов модальностей используется подбор параметров по сетке (Grid search) с процедурой скользящего контроля. Выборка D^N разбивается на K различными способами на две непересекающиеся подвыборки: $D^N = D_k^M \cup D_k^L$, где D_k^M - обучающая выборка, D_k^L - контрольная выборка, $N = M + L$, $k = 1, ..., K$ - номер разбиения.

Для каждого разбиения k строится мультимодальная тематическая модель a_k и вычисляется функционал качества $Q_k = f1_{macro}(a_k, D_k^L)$. Среднее арифметическое значений по всем разбиениям называется оценкой скользящего контроля:

$$CV = \frac{1}{K} \sum_{k=1}^{K} Q(a_k(D_k^M), D_k^L) \tag{5}$$

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