

# Метаобучение тематических моделей классификации

А. С. Ватолин<sup>1</sup>, Ю. А. Сердюк<sup>2</sup>, К. В. Воронцов<sup>1</sup>

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

<sup>1</sup> 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 [?], для классификации [?] и суммаризации текстов [?], а также для ранжирования статей [?]. Одним из продвинутых инструментов, который реализует все из перечисленных выше инструментов, является библиотека BigARTM [?]. 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 [?] 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 [?] 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 [?] 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 [?]. 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

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.  $W^1, \dots, W^m$  - 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  $w_1, \dots, w_{n_d} \in W^m \forall m \in M$ , 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:

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

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

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

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

### 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.