Проблема несбалансированности

Виктор Панкратов

6 марта 2021 г.

1 Постановка задачи

1.1 Общая постановка

Пусть D - множество документов, T - конечное множество тем. Каждый из документов $d \in D$ задается его длиной n_d и последовательностью термов $\{w_i \in W\}_{i=1}^{n_d}$, элементы которой в дальнейшем будем называть словами. Вероятностная модель порождения коллекции вводится при следующих дополнительных предположениях

- Гипотеза мешка слов: вышеописанное представление документа эквивалентно представлению документа в виде неупорядоченного множества входящих в него слов, в которое каждое слово w входит n_{wd} раз.

- Гипотеза о существовании тем: каждое вхождение слова в документ связано с некоторой темой $t \in T$

- Гипотеза условной независимости: вероятность появления слова w в документе d по теме t не зависит от документа d и описывается распределением
$$p(w|d,t) = p(w|t)$$

При таких условиях вероятность появления слова w в документе d описывается распределениями $p(w|t) = \phi_{wt}, p(t|d) = \theta_{td}$. Задача тематического моделирования заключается в нахождении этих распределений. Это эквивалентно задаче получения матричного разложения

$$F = \Phi\Theta \tag{1}$$

$$F = \left(\frac{n_{wd}}{n_d}\right)_{W\times D} \quad \Phi = (\phi_{wt})_{W\times T} \quad \Theta = (\theta_{td})_{T\times D}$$

Данную задачу решают максимизацией с помощью EM-алгоритма функции правдоподобия

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{2}$$

Задача (1) поставлена некорректно: в общем случае ее множество решений бесконечно. Чтобы уменьшить множество решений, в функцию (2)добавляют один или несколько регуляризаторов, зависящих от матриц Φ, Θ. Функция правдоподобия при этом принимает следующий вид:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_{i} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta} \tag{3}$$

1.2 Проблема несбалансированности

Вышеописанная модель склонна выделять равномощные темы, то есть при n_t определенном как $n_t = \sum\limits_{d\in D} p(t|d) n_d \Rrightarrow \forall\ t_i, t_j \in T \to \frac{n_{t_1}}{n_{t_2}} \approx 1$, что получило название "проблема несбалансированности". Такой эффект возникает из-за изначальной постановки задачи: при максимизации правдоподобия модели выгодно использовать все свои параметры.В свою очередь, сокращение долей отдельных тем приводит к неполному использованию, а в пределе - к уменьшению числа параметров. В реальных же коллекциях темы могут оказаться несбалансированными. Чтобы модель и в таком случае корректно выделила темы, в нее предлагается добавить регуляризатор R:

$$R = \sum_{d\in D}\sum_{w\in d} \beta_{dw} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \tag{4}$$

$$\beta_{dw} = \sum_{t\in T} \frac{p(t|d,w)}{p_t} \qquad p_t = \frac{n_t}{n}$$

Введение данного регуляризатора эквивалентно требованию минимизации суммарной семантической неоднородности тем.

$$\sum_{t \in T} S_t = \sum_{d \in D} \sum_{w \in d} \left(\sum_{t \in T} \frac{n_{tdw}}{n_t} \right) \ln \frac{\hat{p}(w|d)}{p(w|d)} \to \min_{\Phi,\Theta} \tag{5}$$

В данной работе будет исследована зависимость качества получаемого решения задачи(1) от добавления регуляризатора R, показано, как введение этого регуляризатора изменяет статистику S_t а также проанализировано влияние R в совокупности с другими регуляризаторами, представленными в работах ранее.

1.2.1 Регуляризатор декоррелирования

$$R_1 = -\sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \tag{6}$$

Введение данного регуляризатора эквивалентно требованию находить различные темы, т е уменьшению ковариации тем.

1.2.2 Регуляризатор сглаживания

$$R_2 = -\sum_{t \in T} \sum_{w \in W} \alpha_{wt} \ln \phi_{wt} \tag{7}$$

Введение данного регуляризатора эквивалентно близости ϕ_t к заданному распределению α_t. Аналогично вводится сглаживание для матрицы Θ

2 Эксперимент

2.1 Генерация коллекции

Для эксперимента будем генерировать синтетическую коллекцию данных. Процесс генерации условно можно разделить на 2 этапа: генерация матриц Φ, Θ и построение документов по ним.

2.1.1 Генерация матриц

Столбцы матриц Φ, Θ порождаются симметричными распределениями Дирихле. Параметр распределения определяется из соображений реалистичности коллекции и берется малым для разреженности получаемых матриц. Для матрицы Φ он берется равным ≈ 0.01, для матрицы $\Theta \approx 0.1$. Чтобы регулировать баланс тем будем на этом этапе менять наибольшие значения в столбцах Θ с необходимыми для эксперимента.

2.1.2 Генерация документов

Для генерации очередного слова w_i сначала генерируется тема t_i документа из соответствующего этому документу столбцу матрицы Θ. Затем слово генерируется из столбца Φ, соответствующего теме t_i. Таким образом, процесс генерации документов описывается как

$$t_i \sim Dir(t|d) \quad w_i \sim Dir(w|t_i) i \in 1 \dots n_d$$

2.2 Стандартная модель

Сгенерируем вышеописанным образом несколько коллекций с различной степенью несбалансированности. Для каждой из них используем стандартную модель для нахождения матриц Φ, Θ. Чтобы оценить сходство полученных матриц Φ_{exp} с используемыми при генерации в данной работе считается количество взаимно ближайших по евклидовой метрике столбцов матриц Φ, Φ_{exp}, то есть пар столбцов $Phi[i]$,$Phi_{exp}[j]$:

$$\arg\min_k \; (dist(Phi[i], Phi_{exp}[k])) = j$$

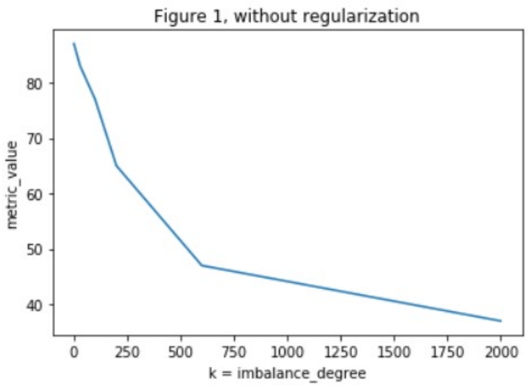
$$\arg\min_k \; (dist(Phi[k], Phi_{exp}[j])) = i$$

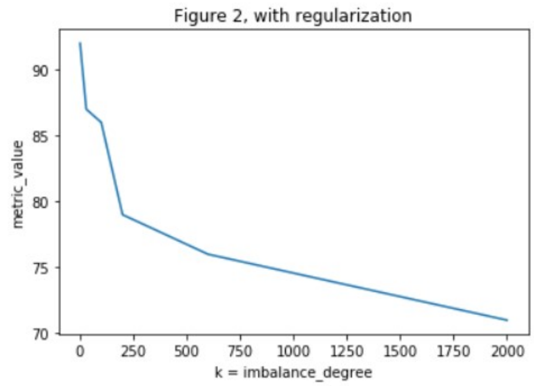
$dist$ в формулах выше - евклидово расстояние. Оно выбрано исключительно для демонстрации в данном отчете.

На графике 1 представлены результаты описанного эксперимента. Видно, что при увеличении степени несбалансированности качество решения падает.

2.3 Добавление регуляризатора

Теперь рассмотрим, как введение регуляризатора R улучшает качество модели. Добавим регуляризатор R в модель и повторим предыдущий эксперимент.Результаты представлены на графике 2.





В данном эксперименте использовался постоянный коэффициент регуляризации $\tau = 0.5$.Видно, что качество решения продолжает ухудшаться при увеличении степени несбалансированности, однако гораздо медленнее, чем в первом случае.