Отбор тем в вероятностных тематических моделях

Плавин Александр Викторович

Студент

Факультет УПМ МФТИ, Москва, Россия

E-mail: alexander@plav.in

Вероятностная тематическая модель коллекции документов описывает каждый документ $d \in D$ дискретным вероятностным распределением $p(t|d)$ на множестве тем, а каждую тему $t \in T$ — распределением $p(w|t)$ на множестве слов $w \in W$. Преобразование коллекции из исходного формата счетчиков n_{dw} слов в документах в две матрицы вероятностных распределений $\Phi_{wt} = p(w|t)$ и $\Theta_{td} = p(t|d)$ оказывается полезным в задачах поиска, классификации, кластеризации, аннотирования документов.

При построении тематической модели важной проблемой является определение числа тем, представленных в коллекции. В случае неправильного нахождения этого значения темы могут смешиваться или наоборот, разбиваться на несколько, теряя интерпретируемость для пользователей модели. Одним из наиболее популярных методов определения числа тем на сегодняшний день является Hierarchical Dirichlet Process (HDP, [3]), однако эта модель определяет число тем неустойчиво, и результат существенно зависит от начального приближения.

Данная работа проведена в рамках альтернативного подхода — аддитивной регуляризации тематических моделей [1, 2], который позволяет вводить в модель дополнительные требования посредством добавления к оптимизируемому функционалу соответствующих критериев-регуляризаторов. В таком случае некорректно поставленная задача максимизации логарифма правдоподобия $L(\Phi, \Theta) = \sum_{d,w} n_{dw} p(w|d)$ заменяется задачей $L + \tau R \to \max$, где $R(\Phi, \Theta)$ — регуляризатор, τ – коэффициент регуляризации.

В работе предлагается регуляризатор оптимизации числа тем, который состоит в максимизации KL-дивергенции между равномерным распределением на темы и получаемым в модели:

$$R = \mathrm{KL}\left(\frac{1}{|T|} \| \sum_{d \in D} p(t|d)p(d)\right)$$

. Он соответствует полному исключению из модели наименее значи-

мых, плохо представленных в коллекции тем.

Обучение модели производится с помощью EM-алгоритма, итеративно просматривающего коллекцию. В качестве начального значения выбирается заведомо избыточное число тем, которое затем уменьшается в ходе итераций и в результате стабилизируется. Также стабилизируется перплексия — количественная оценка качества описания коллекции тематической моделью.

С данным регуляризатором оптимизации числа тем проведена серия экспериментов как на модельных, так и реальных данных. Описанный EM-алгоритм запускался на одних и тех же данных с разными параметрами, и исследовалась зависимость результирующего числа тем от параметров. В результате была выработана методика, позволяющая определять число тем полностью автоматически. В экспериментах с модельными данными, которые генерировались с известным истинным числом тем, предлагаемый подход находит это число с высокой точностью и превосходит известные методы определения числа тем HDP и PTM (Parsimonious Topic Models, [4]). На данных реальных коллекций метод работает устойчиво и почти не зависит от начального приближения, однако диапазон значений, определяемый как оптимальный, оказывается достаточно широким. Это объясняется наличием в реальных данных как меньшего числа более крупных тем, так и большего числа более узких тем.

Дополнительным эффектом предлагаемого регуляризатора является повышение различности тем: в результате остаются преимущественно темы с линейно-независимыми множествами слов, а их линейные комбинации исключаются из модели. Это способствует повышению интерпретируемости выявляемых тем.

Литература

1. Воронцов К. В. Лекции по тематическому моделированию `http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf`
2. Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Analysis of Images, Social Networks, and Texts — CCIS № 436, Springer 2014, P. 29–46.
3. Teh Y. W. et al. Hierarchical Dirichlet Processes // Journal of The American Statistical Association № 101, 2006, P. 1566-1581.
4. Hossein Soleimani, Miller J. D. Parsimonious Topic Models with Salient Word Discovery `http://arxiv.org/abs/1401.6169`