УДК 004.912

Последовательное добавление и удаление тем в вероятностных тематических моделях

А.В. Плавин
Московский физико-технический институт (государственный университет)

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 Φ и Θ 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, [4]), 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, 3], который позволяет вводить в модель дополнительные требования посредством добавления к оптимизируемому функционалу соответствующих критериев-регуляризаторов. При этом некорректно поставленная задача максимизации логарифма правдоподобия $L(\Phi,\Theta) = \sum_{d \in D, w \in W} n_{dw} \sum_{t \in T} p(w|t)\, p(t|d)$ заменяется задачей $L + \tau R(\Phi,\Theta) \to \max_{\Phi,\Theta}$, где $R(\Phi,\Theta)$ – регуляризатор, τ – коэффициент регуляризации.

В работе предлагается использовать регуляризатор отбора тем, который состоит в максимизации $R(\Phi,\Theta) = -|T| = -\sum_{w,t} \Phi_{w,t}$ одновременно с логарифмом правдоподобия. Он соответствует полному исключению из модели наименее значимых, плохо представленных в коллекции тем. Для поэтапного добавления новых тем добавляется их заведомо избыточное количество, обучается модель, и лишние исключаются с помощью аналогичного регуляризатора.

Обучение модели производится с помощью EM-алгоритма, итеративно просматривающего коллекцию. В качестве начального значения выбирается некоторогое число тем, которое затем изменяется 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, [5]). На данных реальных коллекций (статьи конференции NIPS за 12 лет на английском языке и конференций ММРО-ИОИ за 3 года на русском языке) 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.

Алгоритм последовательного удаления и добавления тем показывает лучшие результаты по определению числа тем, чем только отбор тем без добавления новых.

Литература

1. *Vorontsov K. V., Potapenko A.A., Plavin A.V.* Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // Statistical Learning and Data Sciences, Springer, 2015.
2. *Воронцов К.В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. Т. 455, №3.
3. *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.
4. *Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei* Hierarchical Dirichlet Processes // Journal of The American Statistical Association – 2006.
5. *Hossein Soleimani and David J. Miller* Parsimonious Topic Models with Salient Word Discovery.