Аддитивная регуляризация тематических моделей коллекций текстовых документов

К. В. Воронцов (voron@forecsys.ru)

10 января 2014 г.

Тематическое моделирование — одно из активно развивающихся направлений статистического анализа текстов [1]. *Вероятностная тематическая модель* выявляет тематику коллекции текстовых документов, описывая каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотации, суммаризации текстов.

Пусть заданы три конечных множества: коллекция текстовых документов D, словарь терминов W и множество тем T. Предполагается, что порядок терминов в документах не важен, и коллекция является случайной выборкой из дискретного распределения $p(d, w, t)$ на $D \times W \times T$. Переменные d и w являются наблюдаемыми, переменная t — латентной, т. е. появление каждой пары (d, w) связано с некоторой неизвестной темой t. Коллекция представляется матрицей частот $F = (\hat{p}_{wd})_{W \times D}$, где $\hat{p}_{dw} = n_{dw}/n_d$ — частотная оценка условной вероятности $p(w \,|\, d)$, n_{dw} — число вхождений термина w в документ d, n_d — длина документа d.

Модель *вероятностного латентного семантического анализа* PLSA [2] описывает условную вероятность появления терминов в документах

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

через неизвестные условные распределения $p(w \,|\, t) \equiv \varphi_{wt}$ для каждой темы $t \in T$ и $p(t \,|\, d) \equiv \theta_{td}$ для каждого документа $d \in D$. Задача сводится

к поиску стохастического матричного разложения $F = \Phi\Theta$. Для поиска приближённого решения максимизируется логарифм правдоподобия:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \;\to\; \max_{\Phi, \Theta}. \tag{2}$$

Для максимизации (2) применяется EM-алгоритм [2, 3], в котором итерационно повторяются два шага.

На E-шаге по формуле Байеса оцениваются условные распределения латентных тем $p(t \mid d, w)$ для всех терминов в документах (d, w):

$$p(t \mid d, w) = \frac{\varphi_{wt} \theta_{td}}{\sum\limits_{s \in T} \varphi_{ws} \theta_{sd}}. \tag{3}$$

На M-шаге по этим условным вероятностям вычисляются частотные оценки искомых условных вероятностей:

$$\varphi_{wt} \propto \hat{n}_{wt} = \sum_{d \in D} n_{dw} p(t \mid d, w), \qquad \theta_{td} \propto \hat{n}_{dt} = \sum_{w \in d} n_{dw} p(t \mid d, w), \tag{4}$$

где знак пропорциональности $\propto$ означает, что выражение справа необходимо нормировать, чтобы слева получить распределение.

EM-алгоритм хорошо изучен, доказана его сходимость к локальному максимуму правдоподобия. В [4] описаны различные способы реорганизации итераций для повышения скорости сходимости.

Модель *латентного размещения Дирихле* LDA [5] вводит дополнительное вероятностное предположение, что распределения φ_t и θ_d, как вектор-столбцы матриц Φ и Θ, порождаются распределениями Дирихле с гиперпараметрами $\beta = (\beta_w)_{w \in W}$ и $\alpha = (\alpha_t)_{t \in T}$, соответственно, что приводит к сглаживанию частотных оценок на M-шаге:

$$\varphi_{wt} \propto \hat{n}_{wt} + \beta_w, \qquad \theta_{td} \propto \hat{n}_{dt} + \alpha_t. \tag{5}$$

Другие различия между EM-подобными алгоритмами для моделей PLSA и LDA являются второстепенными [3]. Более того, известные их модификации в равной мере применимы к обеим моделям [4].

Модель LDA стала де факто базовой для сотен модификаций, приспособленных под решение самых разных задач. В то же время, LDA порождает две открытые проблемы, мало упоминаемые в литературе.

Во-первых, априорные распределения Дирихле и их обобщения — процессы Дирихле и Питмана-Йора — имеют слабые лингвистические обоснования и не моделируют какие-либо явления естественного языка. Их применение продиктовано исключительно математическим удобством — возможностью аналитического интегрирования по пространству параметров модели при использовании байесовского вывода.

Во-вторых, для практических задач необходимы композитные модели, совмещающие большое число функциональных требований [1]. В частности, для научного поиска по большим коллекциям публикаций нужна модель, одновременно иерархическая, динамическая, n-граммная, разреженная, робастная, мультиязычная, и т. д. Байесовский вывод оказывается слишком громоздким для совмещения более 2–3 требований в одной модели. Такие модели в литературе пока не рассматривались.

Таким образом, назрела необходимость в разработке новых принципов построения тематических моделей, свободных от избыточных вероятностных допущений и упрощающих построение композитных моделей. Предлагаемая теория *аддитивной регуляризации тематических моделей* (АРТМ) решает эти проблемы.

Стохастическое матричное разложение $\Phi\Theta$ не единственно и определено с точностью до невырожденного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$. Таким образом, задача построения тематической модели является некорректно поставленной, и для её решения необходимо применять регуляризацию. Вместо байесовской регуляризации предлагается опираться на более общую концепцию тихоновской регуляризации [6].

Допустим, что наряду с правдоподобием (2) требуется максимизировать ещё n критериев $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$, называемых *регуляризаторами*. Для решения задачи многокритериальной оптимизации будем максимизировать линейную комбинацию критериев L и R_i с неотрицательными *коэффициентами регуляризации* τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta), \qquad L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}. \qquad (6)$$

Данная задача по-прежнему решается с помощью EM-алгоритма,

но с модифицированной формулой М-шага:

$$\varphi_{wt} \propto \left(\hat{n}_{wt} + \varphi_{wt}\frac{\partial R(\Phi,\Theta)}{\partial \varphi_{wt}}\right)_+, \qquad \theta_{td} \propto \left(\hat{n}_{dt} + \theta_{td}\frac{\partial R(\Phi,\Theta)}{\partial \theta_{td}}\right)_+. \quad (7)$$

Добавление ещё одного регуляризатора приводит к добавлению соответствующей поправки в формуле М-шага. Именно это и позволяет строить *композитные тематические модели*, сочетающие в себе многие дополнительные требования, в том числе не вероятностного характера.

Далее приводятся примеры регуляризаторов, как известных в литературе (хотя не всегда очевидно, что это регуляризатор), так и новых. Список регуляризаторов ни в коей мере не претендует на полноту и носит скорее обзорно-иллюстративный характер.

1. *Сглаживающий регуляризатор* формализует требование, чтобы распределения φ_t и θ_d были близки к заданным дискретным распределениям $\tilde{\beta}$ и $\tilde{\alpha}$ по дивергенции Кульбака–Лейблера:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{w\in W} \tilde{\beta}_w \ln \varphi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in T} \tilde{\alpha}_t \ln \theta_{td} \to \max,$$

где β_0, α_0 — коэффициенты регуляризации. Дифференцирование R немедленно приводит к формулам (5) для М-шага в LDA, если ввести обозначения $\beta_w = \beta_0\tilde{\beta}_w$, $\alpha_t = \alpha_0\tilde{\alpha}_t$. При этом не используются ни априорные распределения Дирихле, ни байесовский вывод.

В теории АРТМ распределение Дирихле утрачивает центральную роль. Это лишь один из возможных регуляризаторов, не самый лучший и не настолько универсальный, как принято считать. В качестве базовой модели логичнее брать PLSA, не имеющую собственных регуляризаторов, и добавлять к ней проблемно-ориентированные регуляризаторы.

2. *Разреживающий регуляризатор.* Естественно предположить, что каждый документ и каждый термин связан с небольшим числом тем. Тогда среди вероятностей φ_{wt} и θ_{td} должно быть много нулевых. Это противоречит модели LDA, так как распределение Дирихле не допускает нулевых значений в порождаемых им векторах.

Чем сильнее разрежено распределение, тем ниже его энтропия. Максимальной энтропией обладает равномерное распределение. Поэтому

возьмём регуляризатор, максимизирующий дивергенцию между равномерным распределением и искомыми распределениями:

$$R(\Phi,\Theta) = -\beta \sum_{t\in T}\sum_{w\in W} \ln \varphi_{wt} - \alpha \sum_{d\in D}\sum_{t\in T} \ln \theta_{td} \to \max.$$

Получим формулу М-шага, которая отличается от сглаживающего регуляризатора знаком параметра и приводит к разреживанию:

$$\varphi_{wt} \propto \big(\hat{n}_{wt} - \beta\big)_+, \qquad \theta_{td} \propto \big(\hat{n}_{dt} - \alpha\big)_+.$$

3. *Регуляризатор для частичного обучения.* Для улучшения интерпретируемости тематической модели эксперты могут задавать обучающие данные. Пусть для части документов $d \in D_0$ известно, что они относятся к темам $T_d \subset T$, для части тем $t \in T_0$ известно, что к ним относятся термины $W_t \subset W$. Возьмём распределение φ^0_{wt}, равномерное на W_t, и распределение θ^0_{td}, равномерное на T_d. Рассмотрим регуляризатор

$$R(\Phi,\Theta) = \tau_1 \sum_{t\in T_0}\sum_{w\in W_t} \varphi^0_{wt} \ln \varphi_{wt} + \tau_2 \sum_{d\in D_0}\sum_{t\in T_d} \theta^0_{td} \ln \theta_{td} \to \max.$$

Формулы М-шага, согласно (7), принимают вид

$$\theta_{td} \propto \hat{n}_{dt} + \tau_1\theta^0_{td}, \quad d \in D_0; \qquad \varphi_{wt} \propto \hat{n}_{wt} + \tau_2\varphi^0_{wt}, \quad t \in T_0.$$

Это тоже сглаживание, но, в отличие от LDA, оно производится только для тех θ_{td} и φ_{wt}, по которым имеются обучающие данные.

4. *Ковариационный регуляризатор для тем.* Считается, что повышение различности тем улучшает интерпретируемость модели [7]. Регуляризатор, минимизирующий ковариации между вектор-столбцами φ_t,

$$R(\Phi) = -\frac{\tau}{2}\sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \varphi_{wt}\varphi_{ws} \to \max,$$

приводит к формуле М-шага

$$\varphi_{wt} \propto \Big(\hat{n}_{wt} - \tau\varphi_{wt}\sum_{s\in T\setminus t} \varphi_{ws}\Big)_+.$$

Смысл этой формулы в том, что условные вероятности $\varphi_{wt} = p(w\,|\,t)$ постепенно уменьшаются для тех слов w, которые имеют бóльшие значения вероятности φ_{ws} в других темах. В процессе итераций EM-алгоритма для каждого слова вероятности наиболее значимых тем приобретают всё большие значения, а вероятности менее значимых тем уменьшаются и могут обращаться в нуль. Таким образом, данный регуляризатор также является разреживающим. Кроме того, он обладает дополнительным полезным свойством группировать стоп-слова в отдельные темы [7].

5. *Ковариационный регуляризатор для документов.* Иногда имеется дополнительная информация о связях между документами схожей тематики. В частности, это могут быть документы, относящиеся к одной рубрике тематического рубрикатора, отложенные пользователем электронной библиотеки в одну папку, или ссылающиеся друг на друга. Пусть задан ориентированный граф $G = \langle D, E\rangle$, и рёбра графа $(d, c) \in E$ означают, что тематика документа c близка к тематике документа d. Формализуем это предположение с помощью регуляризатора:

$$R(\Theta) = \tau \sum_{(d,c)\in E} n_{dc} \sum_{t\in T} \theta_{td}\theta_{tc} \to \max,$$

где n_{dc} — вес ребра (d, c), например, число ссылок на c из d. В [8] предложена похожая модель LDA-JS, в которой вместо максимизации ковариации минимизируется дивергенция Йенсена-Шеннона между θ_d и θ_c. Формула M-шага для θ_{td}, согласно (7), принимает вид

$$\theta_{td} \propto \hat{n}_{dt} + \tau\theta_{td} \sum_{c:\ (d,c)\in E} n_{dc}\theta_{tc}.$$

Таким образом, условные распределения $\theta_{td} = p(t\,|\,d)$ в ходе итераций приближаются к распределениям θ_{tc} документов, связанных с d.

6. *Максимизация когерентности.* Тема называется *когерентной*, если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции. Средняя когерентность тем считается хорошей мерой интерпретируемости тематической модели. Пусть C_{uv} — оценка совместной встречаемости слов $(u, v) \in Q \subset W^2$. В [9] предлагается формула M-шага для алгоритма сэмплирования Гибб-

са с обоснованием через обобщённую урновую схему Пойя:

$$\varphi_{wt} \propto \hat{n}_{wt} + \tau \sum_{u \in W \setminus w} C_{uw} \hat{n}_{ut}.$$

Нетрудно показать, что эта формула также следует из регуляризатора

$$R(\Phi) = \tau \sum_{t \in T} \sum_{(u,v) \in Q} C_{uv} \hat{n}_{ut} \ln \varphi_{vt} \to \max,$$

минимизирующего сумму дивергенций между каждым распределением φ_{vt} и его эмпирической оценкой по всем словам, встречающимся с v.

7. *Максимизация правдоподобия в задачах классификации.* Пусть имеется дополнительная информация о классификации документов, и предполагается, что документы одного класса часто имеют схожую тематику. В роли классов могут выступать: категории рубрикатора, авторы, годы публикаций, цитирующие или цитируемые авторы или документы, пользователи (читатели) документов. Для всех этих случаев были разработаны специальные модели [1]. Пусть каждому документу d соответствует набор элементов C_d из конечного множества C. Задача заключается в том, чтобы выявить связи между классами и темами, улучшить качество тематической модели с помощью дополнительной информации о классификациях, построить алгоритм классификации новых документов. Одна из лучших тематических моделей классификации Dependency LDA [10] определяет распределение классов документов $p(c \,|\, d)$ через распределения классов тем $\psi_{ct} = p(c \,|\, t)$ и тем документов $\theta_{td} = p(t \,|\, d)$ по аналогии с основной тематической моделью (1):

$$p(c \,|\, d) = \sum_{t \in T} \psi_{ct} \theta_{td}, \tag{8}$$

где новой неизвестной является *матрица классов тем* $\Psi = (\psi_{ct})_{C \times T}$. В оригинальной работе приводится довольно громоздкий вывод алгоритма сэмплирования Гиббса в рамках байесовского подхода. Однако ровно тот же результат достигается с помощью регуляризатора, минимизирующего дивергенцию между моделью классификации $p(c \,|\, d)$ и эмпирической частотой классов в документах m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td} \to \max, \tag{9}$$

где коэффициент регуляризации τ необходим для «приведения к одному масштабу» частот слов n_{dw} и частот классов m_{dc}.

При линейном комбинировании регуляризаторов R_i возникает проблема выбора вектора коэффициентов $\tau = (\tau_i)_{i=1}^n$. Аналогичная проблема эффективно решается в эластичных сетях (elastic net) при комбинировании L_1- и L_2-регуляризации для задач регрессии и классификации [11]. В задачах тематического моделирования разнообразие регуляризаторов гораздо больше, и они нетривиальным образом влияют друг на друга. Предварительные эксперименты показали, что некоторые регуляризаторы могут ухудшать сходимость, если включать их слишком рано или слишком резко. Поэтому предлагается увеличивать коэффициенты регуляризации постепенно, по определённой траектории в пространстве коэффициентов регуляризации. Построение таких траекторий в задачах тематического моделирования пока является открытым вопросом.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект №11-07-00480) и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Список литературы

[1] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. — 2010. — Vol. 4, no. 2. — Pp. 280–301.

[2] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.

[3] *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada. — 2009. — Pp. 27–34.

[4] *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. — 2012. — Т. 4, № 4. — С. 693–706.

[5] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — Vol. 3. — Pp. 993–1022.

[6] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1986.

[7] *Tan Y., Ou Z.* Topic-weak-correlated latent dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP), China, Tainan. — 2010. — Pp. 224–228.

[8] *Dietz L., Bickel S., Scheffer T.* Unsupervised prediction of citation influences // Proceedings of the 24th international conference on Machine learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 233–240.

[9] *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 262–272.

[10] *Rubin T. N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // Machine Learning. — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.

[11] *Friedman J. H., Hastie T., Tibshirani R.* Regularization paths for generalized linear models via coordinate descent // Journal of Statistical Software. — 2010. — Vol. 33, no. 1. — Pp. 1–22.