01

Оптимизация метода семплирования по Гиббсу для анализа гранулированной среды

© *С.Н. Кольцов, С.И. Николенко, Е.Ю. Кольцова*

Национальный исследовательский университет „Высшая школа экономики“ (НИУ ВШЭ), Санкт-Петербург
E-mail: skoltsov@hse.ru

Поступило в Редакцию 26 февраля 2016 г.

Предлагается новая вариация метода восстановления плотности распределений вероятностей для задач тематического моделирования. Рассматриваются недостатки алгоритма сэмплирования по Гиббсу и предлагается его модифицированный вариант — гранулированный метод сэмплирования. На основе статистического моделирования показано, что предлагаемый алгоритм является более стабильным по сравнению с двумя другими вариантами алгоритма сэмплирования.

В связи с бурным развитием вычислительной техники наблюдается резкий рост в области методов статистического моделирования, в особенности в моделях с применением марковских цепей (MCMC-моделирование). Байесовский подход к оценке скрытых параметров многомерных распределений де-факто стал одним из ведущих методов в таких направлениях, как физика высоких энергий [1], масс-спектрометрия и биоинформатика [2], астрофизика [3], статистическая физика [4] и др. Одним из наиболее распространенных алгоритмов оценки скрытых параметров на основе наблюдаемых данных является алгоритм сэмплирования по Гиббсу. Алгоритм примечателен тем, что для него не требуется явно выраженное совместное распределение, а нужны лишь одномерные условные вероятности, входящие в распределение.

В последнее время методы, разрабатываемые в физике, все чаще применяются для анализа больших данных (big data) в области data mining, которые характеризуются многомерными распределениями. Несмотря на то что big data может представлять собой набор различных объектов, таких как масс-спектры [2], звуковые дорожки или фотографии [5], а также новостные тексты в социальных сетях [6],

задача восстановления исходной плотности распределений одинакова для таких разнообразных объектов.

Задача восстановления исходного многомерного распределения в виде смеси распределений со скрытыми параметрами называется тематическим моделированием (ТМ) [7,8]. ТМ основано на следующих положениях: 1. Оценка апостериорного распределения на основе правила Байеса находится при помощи оценки математического ожидания через сэмплирование. 2. Распределения тем в документах и слов в темах — это мультиноминальные распределения с априорными распределениями Дирихле с параметрами α и β. Подход к оценке плотности распределения в тематическом моделировании с учетом функций Дирихле называется латентным размещением Дирихле (Latent Dirichlet allocation, LDA) [7]. В данной работе рассматривается модификация метода сэмплирования по Гиббсу для текстовых данных [8], который уже применялся для мягкой кластеризации масс-спектров [2], обнаружения и идентификации ядерных изотопов [9] и во многих других приложениях.

В рамках тематического моделирования на текстовых данных наблюдаемыми переменными являются документы d и слова w из заданной коллекции. Также предполагается, что существует конечное множество тем T и коллекция документов порождается дискретным распределением $p(d, w, t)$, где d — документ, w — слово, t — тема. Переменная t, характеризующая счетное множество тем, является скрытой. Под темой t понимается одномерное распределение Дирихле на словах. Соответственно каждый документ представлен смесью распределения латентных тем, а каждая тема определяется вероятностным распределением на множестве слов. Построить тематическую модель данных означает найти скрытые распределения слов и документов по темам на основе наблюдаемых переменных, т. е. множество одномерных условных распределений $p(w|t) \equiv \varphi(w, t)$ (матрица Φ, распределение слов по темам) и множество одномерных распределений $p(t|d) \equiv \theta(t, d)$ (матрица Θ, распределение документов по темам) для каждого документа d. Итоговые распределения слов и документов на основе сэмплирования Гиббса рассчитываются по следующим формулам [8]:

$$P(z_i = j|w_i = m, z_{-i}, w_{-i}) \approx \frac{C^{WT}_{m,j} + \beta}{\sum\limits_{\dot{m}} C^{WT}_{\dot{m},j} + V\beta} \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j} + \alpha T}, \qquad (1)$$

$$\theta_{dj} = \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j} + T\alpha}, \qquad (2)$$

$$\phi_{m,j} = \frac{C^{WT}_{m,j} + \beta}{\sum\limits_{\dot{m}} C^{WT}_{\dot{m},j} + V\beta}, \tag{3}$$

где α, β — параметры, определяющие одномерные распределения Дирихле; C — счетчики, получаемые в ходе сэмплирования, а именно: $C^{WT}_{m,j}$ — число раз, которые слово w встречалось в теме t; $C^{DT}_{d,j}$ — число раз, которые слово w в документе d было связано с темой t; $\Sigma_m C^{WT}_{mj} = n_t$ — число слов, связанных с темой t; $C^{DT}_{d,j} = n_d$ — длина документа в словах. В ходе процедуры сэмплирования производится расчет матрицы $p(z = j|w_i = m, z_{-i}, w_{-i})$. На основании этой матрицы производится расчет счетчиков C по формуле (1) и затем уже при помощи счетчиков производится расчет итоговых распределений $\theta_{d,j}$ и $\phi_{m,j}$ по формулам (2), (3).

Решение задачи тематического моделирования эквивалентно стохастическому матричному разложению, в котором большая матрица F, содержащая документы d и слова w, аппроксимируется произведением двух матриц $\theta_{d,j}$ и $\phi_{m,j}$ меньшей размерности. Однако стохастическое матричное разложение определено не единственным образом, а с точностью до невырожденного преобразования [10]. В терминах алгоритма сэмплирования по Гиббсу неоднозначность восстановления многомерной плотности смеси распределений связана с тем, что алгоритм, стартуя из различных начальных приближений, будет сходиться к различным точкам из множества решений. Это выражается в том, что при разных запусках алгоритма на одних и тех же исходных данных содержимое матриц $\theta_{d,j}$ и $\phi_{m,j}$ будет различным, т. е. алгоритм сэмплирования Гиббса не является стабильным. Задачи, решение которых не единственно или неустойчиво, называются некорректно поставленными. Общий подход к их решению дает регуляризация по Тихонову [11], которая заключается в доопределении априорной информации, что позволяет сузить множество решений.

В настоящей работе предлагается метод гранулированного сэмплирования (Granulated LDA), который отличается от обычного метода сэмплирования по Гиббсу [8] модификацией понятия темы. Во-первых, каждый документ рассматривается как гранулированная поверхность, состоящая из гранул. Под гранулой понимается последовательность слов заданной длины. Во-вторых, все слова, находящиеся в одной грануле, относятся к одной теме. Гранулы характеризуются своим размером — шириной окна сэмплирования (регуляризационный параметр). Целью гранулированного варианта ТМ является восстановление

матриц $\theta_{d,j}$ и $\phi_{m,j}$ путем усреднения тематического содержимого гранул по большому числу документов, т. е. расчетом математического ожидания.

Гранулированный вариант сэмплирования реализован следующим образом. После инициализации матриц $\theta_{d,j}$ и $\phi_{m,j}$ организованы два вложенных цикла. Первый цикл — внешний, пробегает по списку документов. Во втором, внутреннем цикле реализовано случайное сэмпилирование по гранулам. В ходе него все слова внутри случайно выбранного окна присваиваются одной теме, номер которой также генерируется случайным образом. Число случайных „сэмплов" слов в документе равно числу слов в документе. В ходе длительного сэмплирования слова, которые часто находятся внутри гранулы, будут чаще иметь одинаковый номер темы, и их вероятность принадлежать одной теме в среднем будет выше. На последнем этапе, на основании счетчиков производится окончательный расчет матриц распределений слов и документов по темам $\theta_{d,j}$ и $\phi_{m,j}$.

Стабильность ТМ в данной работе определялась при помощи сравнения серии одномерных распределений (тем) из разных запусков друг с другом на основе нормализированной меры Кульбака–Лейблера (Kn) [12]. Как показали исследования, две темы являются идентичными, если величина $Kn > 90\%$ [12]. В данном исследовании тема считается стабильно воспроизводящейся от запуска ТМ к запуску, если она воспроизводится в трех запусках на уровне $Kn > 90\%$. Для оценки стабильности тематического моделирования были проведены исследования стабильности трех тематических моделей (на основе сэмплирования Гиббса): 1) LDA (стандартный вариант), 2) SLDA (модель с обучением) [6], 3) GLDA (Granulated LDA, размер гранулы $+1$).

Каждая из трех моделей запускалась по три раза на одних и тех же исходных данных. В тестировании моделей были использованы документы из социальной сети „Живой журнал" общим количеством 101481 документ. В моделировании было использовано 200 тем, параметры функций Дирихле для всех моделей были приняты равными $\alpha = 0.1$ и $\beta = 0.5$. Результаты моделирования показали, что стандартная модель LDA обеспечивает 74 стабильные темы из 200, модель SLDA обеспечивает 87 стабильных тем, а модель GLDA обеспечивает 195 стабильных тем.

Таким образом, статистическое моделирование показало, что предложенная гранулированная модель (GLDA) существенно превосходит

по стабильности аналогичные модели и может эффективно использоваться при решении задач, связанных с проблемой восстановления многомерных распределений, как в физике, так и в других областях (например, data mining), в которых используются статистические физико-математические модели.

Список литературы

[1] *Caldwell A., Kollar D., Kröninger K.* // Comp. Phys. Comm. 2009. V. 180. P. 2197–2209; arXiv:0808.2552.
[2] *Chernyavsky I., Alexandrov T., Maass P., Nikolenko S.* // German Conference on Bioinformatics. 2012. September, P. 39–48.
[3] Handbook of Markov Chain Monte Carlo (Eds. S. Brooks, A. Gelman, G. Jones, X.-L. Meng). Chapman & Hall/CRC Press, 2011. P. 383–399.
[4] *Berg Bernd A., Billoire A.* Markov Chain Monte Carlo Simulations. John Wiley & Sons, Inc., 2008.
[5] *Geman S., Geman D.* // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984. V. 6. P. 721–741.
[6] *Bodrunova S., Koltsov S., Koltsova O., Nikolenko S., Shimorina A.* // Proc. 12th Mexican International Conference on Artificial Intelligence (MICAI 2013). Part I. Berlin: Springer Verlag, 2013. P. 265–274.
[7] *Blei D., Ng A., Jordan M., Lafferty J.* // JMLR. 2003. V. 3. P. 993–1022.
[8] *Griffiths T., Steyvers M.* // Proc. National Academy of Sciences. 2004. V. 101 (Suppl. 1). P. 5228–5335.
[9] *Nelson Craig* et al. // IEEE Conference on Technologies for. IEEE, 2012.
[10] *Vorontsov K.* // Doklady Akademii Nauk. 2014. V. 456. N 3. P. 268–271.
[11] *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. М.: Наука, 1986.
[12] *Koltsov S., Koltsova O., Nikolenko S.* // Proceedings of WebSci'14 ACM Web Science Conference. Bloomington, IN, USA. June 23–26, 2014. NY: ACM, 2014. P. 161–165.