Обобщенные модальности в вероятностных тематических моделях для транзакционных данных

Хрыльченко Кирилл Ярославович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: elightelol@gmail.com

Научный руководитель *— Воронцов Константин Вячеславович*

Транзакционные данные, которые с точки зрения реляционной алгебры содержат информацию в кортежах вида *(продавец, покупатель, дата, сумма транзакции, текст платежного поручения)*, позволяют выявлять профили экономической деятельности для юридических лиц, а также потребительские профили для физических лиц. Основным инструментом моделирования является тематическое моделирование — математическая модель, которая формализует понятие документа $d \in D$, слова в документе $w \in d$ и темы $t \in T$, порождающей слово w в документе d, за счет введения гипотезы условной независимости: $p(w|t,d) = p(w|d)$.

Такая модель не является достаточно сложной для моделирования разнородной информации. Если документом в коллекции является статья, то примером разнородной информации могут выступать текст статьи и список цитирований, которые некорректно моделировать в одном распределении $p(w|t)$. Решением данной проблемы выступает мультимодальное тематическое моделирование, которое позволяет моделировать разнородные источники информции путем введения модальностей, которые задаются распределениями $p(w|t)$ и словарями $W_m, m \in M$. Для такой модели подбор оптимальных параметров осуществляется максимизацией взвешенного правдоподобия по всем модальностям, причем веса модальностей задаются исключительно эмпирически.

Для транзакционных данных юридических лиц документом выступает фирма, а в качестве модальностей используются следующие характеристики: товарные слова платежных поручений, в которых фирма выступает как продавец; товарные слова платежных поручений, в которых фирма выступает как покупатель; контрагенты-покупатели; контрагенты-продавцы и сегмент бизнеса фирмы-документа.

Данная работа исследует влияние весов модальностей на результат моделирования, во многом опираясь на результат теоремы о мультимодальном разложении:

Теорема 1. *Тематическое векторное представление документа $d \in D$ представляется в виде выпуклой комбинации модальных тематических представлений:*

$$\theta_{td} = \sum_{m \in M} \frac{\lambda_m n_d^m}{\sum_{\hat{m} \in M} \lambda_{\hat{m}} n_d^{\hat{m}}} \theta_{td}^m, \quad t \in T,\, d \in D, \tag{1}$$

где $\theta_{td} = p(t|d)$, M — множество модальностей, n_d^m — мощность модальности m в документе d, $(\theta_{td}^m)_{t \in T}$ — тематическое представление документа d для модальности m, λ_m — вес модальности m.

Теорема 1 позволяет формализовать понятие равного вклада модальностей в векторные представления документов, а также явным образом задать веса $\lambda_m, m \in M$, обеспечивающие равный вклад.

Другим существенным следствием теоремы 1 является возможность подбора весов модальностей, оптимизирующих вспомогательный критерий, путем решения на каждом шаге EM-алгоритма задачи вида $J(\Theta) \to \min_\lambda$. Такая возможность исследуется в данной работе на примере задач классификации, в которых объектами классификации являются документы исходной коллекции.

Еще одним результатом данной работы является введение понятия вещественных модальностей — например, модальность логарифмов сумм транзакций, в которых фирма выступает как продавец. В вещественной модальности каждая тема отождествляется с гауссианой, то есть отвечает своим параметрам (μ_t, σ_t^2), задающим нормальное распределение. Тогда тематическое представление документа d позволяет получить смесь нормальных распределений, аппроксимирующую гистограмму логарифмов сумм транзакций документа d. В данной работе также проводится формализация понятия вещественной модальности и вывод необходимой модификации EM-алгоритма для подбора параметров тематической модели.

Литература

1. Hoffman T. Probabilistic latent semantic indexing // 22nd ACM SIGIR conference, 1999, P. 50–57.
2. Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. Т. 455, № 3.