Incremental PLSA

Владимир Герасимов, 4 курс МГУ, ВМК

25 октября 2009 г.

Постановка задачи

U - множество пользователей (users)
R - множество ресурсов (items)
Y - пространство описания транзакций
Протокол транзакций $D = (u_i, r_i, y_i)_{i=1}^{l} \subset U \times R \times Y$ - исходные данные
Исходные данные агрегируются в матрицу кросс-табуляции:

$$F = \|f_{ur}\|_{U \times R}$$

где $f_{ur} = aggr\{(u_i, r_i, y_i) \in D \mid u_i = u,\ r_i = r\}$

Основные задачи

- Прогнозирование незаполненных ячеек $\|f_{ur}\|$
- Оценивание функции сходства $K(u, u'),\ ,K(r, r'),\ K(u, r)$
- Выявление содержательно интерпретируемых латентных характеристик(профилей) как ресурсов, так и клиентов

Алгоритмы коллаборативной фильтрации

Анамнестические алгоритмы семейство алгоритмов которые основаны на хранении всей матрицы кросс-табуляции F и непосредственном поиске в ней схожих клиентов (строк) и ресурсов (столбцов).

Модельные алгоритмы алгоритмы работа которых основана на сравнении векторных характеристик (профилей) клиентов и ресурсов, т.е не требуется вся матрица кросс-табуляции или протокол транзакций.

Обозначения

T - множество всех возможных тем (латентная переменная)

p_{tu} - степень важности темы t для пользователя u
$(p_{tu})_{t \in T}$ - неизвестный профиль клиента u
$P = (p_{tu})_{|T| \times |U|}$ - матрица профилей всех пользователей

q_{tr} - способность ресурса r удовлетворить интерес к теме t
$(q_{tr})_{t \in T}$ - неизвестный профиль ресурса r
$Q = (q_{tr})_{|T| \times |R|}$ - матрица профилей всех ресурсов

λ_t - степень важности темы t

Поиск неизвестных профилей:

За значение элементов матрицы кросс-табуляции естественно принять:

$$\hat{f_{ur}} = \sum_{t \in T} \lambda_t p_{tu} q_{rt}$$

или в матричном виде:

$$\hat{F} = P \Lambda Q^T$$

Для поиска P и Q используется метод наименьших квадратов:

$$\sum_{u \in U} \sum_{r \in R} (\hat{f_{ur}} - f_{ur})^2 = \| F - P \Lambda Q^T \| \to \min_{P, \Lambda, Q}$$

Вероятностный латентный семантический анализ

Основан на явной вероятностной интерпретации величин p_{tu} и q_{rt} как соответствующих условных вероятностей

Профилем клиента назовем вектор условных вероятностей $p_{tu} = p(t|u)$ того, что клиент $u \in U$ интересуется темой $t \in T$, причем должно быть выполнено условие нормировки $\sum_{t \in T} p_{tu} = 1$

Профилем ресурса назовем вектор условных вероятностей $q_{tr} = q(t|r)$ того, что ресурс $r \in R$ удовлетворяет интерес к теме $t \in T$, причем также должно быть выполнено условие нормировки $\sum_{t \in T} q_{tr} = 1$

Вероятностная модель

$$p(u,r) = \sum_{t \in T} p(t)p(u|t)q(r|t)$$

$p(t)$ - вероятность характеризующая интерес к теме $t \in T$
$p(u|t)$ - апостериорное распределение клиентов по теме t
$q(r|t)$ - апостериорное распределение ресурсов по теме t

Принцип максимума правдоподобия:

$$L = \ln \prod_{u \in U} \prod_{r \in R} p(u,r)^{f_{ur}} = \sum_{u \in U} \sum_{r \in R} f_{ur} \ln p(u,r) \to \max_{p(t),p(u|t),q(r|t)}$$

где f_{ur} - значения элементов матрицы кросс-табуляции, количество посещений клиентом u ресурса r.

EM-алгоритм

E-шаг:

$$p(t|u,r) = \frac{p(t)p(u|t)q(r|t)}{\sum\limits_{t' \in T} p(t')p(u|t')q(r|t')}, \quad u \in U, r \in R, t \in T$$

M-шаг:

$$p(t) = \frac{\sum\limits_{u \in U} \sum\limits_{r \in R} f_{ur} p(t|u,r)}{\sum\limits_{u \in U} \sum\limits_{r \in R} f_{ur}}, \quad t \in T;$$

$$q(r|t) = \frac{\sum\limits_{u \in U} f_{ur} p(t|u,r)}{\sum\limits_{u \in U} \sum\limits_{r' \in R} f_{ur'} p(t|u,r')}, \quad r \in R, t \in T$$

$$p(u|t) = \frac{\sum\limits_{r \in R} f_{ur} p(t|u,r)}{\sum\limits_{u' \in U} \sum\limits_{r \in R} f_{u'r} p(t|u',r)}, \quad u \in U, t \in T$$

Замечания:

Профили как условные вероятности тем считаются по формуле Байеса:

- $$p_{tu} = p(t|u) = \frac{p(t)p(u|t)}{\sum_{t' \in T} p(t')p(u|t')} \quad u \in U, t \in T$$
- $$q_{tr} = q(t|r) = \frac{p(t)q(r|t)}{\sum_{t' \in T} p(t')q(r|t')} \quad r \in R, t \in T$$

Важный факт:

- $$p(u,r) = \sum_{t \in T} p(t)p(u|t)q(r|t) =$$
$$= q(r) \sum_{t \in T} p(u|t)q(t|r) = p(u) \sum_{t \in T} q(r|t)p(t|u)$$

Понятие инкрементальности

Под инкрементальностью понимается способность алгоритма расширять и изменять профили в процессе появления новых данных, клиентов и/или ресурсов.
Рассмотрим несколько вариаций PLSA для случая инкрементальности:

- алгоритм Fold-in
- Native IPLSA

Другой вид EM-алгоритма

E-шаг:

$$p(t|u,r) = \frac{p(u|t)q(t|r)}{\sum\limits_{t' \in T} p(u|t')q(t'|r)}, \quad u \in U, r \in R, t \in T \qquad (1)$$

M-шаг:

$$p(u|t) = \frac{\sum\limits_{r \in R} f_{ur} p(t|u,r)}{\sum\limits_{u' \in U} \sum\limits_{r \in R} f_{u'r} p(t|u',r)}, \quad u \in U, t \in T$$

$$q(t|r) = \frac{\sum\limits_{u \in U} f_{ur} p(t|u,r)}{\sum\limits_{u \in U} f_{ur}}, \quad r \in R, t \in T \qquad (2)$$

Fold-in:

При добавлении нового ресурса условные вероятности $q(t|r_{new})$ и $p(t|u, r_{new})$, где r_{new} новый ресурс, оцениваются, с использованием вероятностей $p(u|t)$:

E-шаг:

$$p(t|u, r_{new}) = \frac{p(u|t)q(t|r_{new})}{\sum_{t' \in T} p(u|t')q(t'|r_{new})}, \quad u \in U, t \in T$$

M-шаг:

$$q(t|r_{new}) = \frac{\sum_{u \in U_{r_{new}}} f_{ur_{new}} p(t|u, r_{new})}{\sum_{u \in U_{r_{new}}} f_{ur}}, \quad t \in T$$

$U_{r_{new}}$ - множество клиентов посетивших новый ресурс.
На первой итерации $q(t|r_{new})$ выбирается случайно и нормируется.

Недостатки:

- Новые данные не оказывают влияния на старые
- Изменяются только те параметры профилей, которые относятся к новым данным

Применение этого алгоритма малоэффективно в реальных ситуациях.

Стандартный алгоритм IPLSA

Алгоритм расширения профилей состоит из трех шагов, на каждом из которых для вычисления соответствующих вероятностей, запускается EM-алгоритм. Он работает пока значения вероятностей не стабилизируются.

1й-шаг На первом шаге добавляется новый ресурс r_{new}, вероятности $q(t|r_new)$ определяются случайным образом. $p(u|t)$ на этом шаге не меняется.

E-шаг:

$$p(t|u, r_{new}) = \frac{p(u|t)q(t|r_{new})}{\sum_{t' \in T} p(u|t')q(t'|r_{new})}, \quad u \in U, t \in T$$

M-шаг:

$$q(t|r_{new}) = \frac{\sum_{u \in U_{r_{new}}} f_{ur_{new}} p(t|u, r_{new})}{\sum_{t' \in T} \sum_{u \in U_{r_{new}}} f_{ur_{new}} p(t'|u, r_{new})}, \quad t \in T$$

$$q(r_{new}|t) = \frac{\sum_{u \in U_{r_{new}}} f_{ur_{new}} p(t|u, r_{new})}{\sum_{u \in U_{r_{new}}} \sum_{r \in R + r_{new}} f_{ur} p(t|u, r)}, \quad t \in T$$

2й-шаг На втором шаге изменяются профили клиентов пользовавшихся новыми ресурсами.

E-шаг:

$$p(t|u, r_{new}) = \frac{p(t|u)q(r_{new}|t)}{\sum_{t' \in T} p(t'|u)q(r_{new}|t')}, \quad u \in U_{r_{new}}, t \in T$$

M-шаг:

$$p(t|u) = \frac{f_{ur}p(t|u, r)}{f_{u,r_{new}}}, \quad u \in U_{r_{new}}, t \in T$$

3й-шаг На третьем шаге нормализуется вероятность $p(u|t)$:

$$p(u|t) = \frac{\sum_{r \in (R+r_{new})} f_{ur} p(t|u,r)}{\sum_{r \in (R+r_{new})} \sum_{u' \in U} f_{u'r} p(t|u',r)}, \quad u \in U, t \in T$$

После этого для корректировки запускается алгоритм описанный формулами ??-??

- Лексин В.А. "Двухступенчатые моделии и проблема переобучения в латентном семантическом анализе".
- Hu Wu, Yongji Wang, Xiang Cheng "Incremental Probabilistic Latent Semantic Analysis for Automatic Question Recommendation ACM Conference On Recommender Systems, Proceedings of the 2008 ACM conference on Recommender systems
- T.C.Chouand,M.CChen "Using Incremental PLSA for Threshold Resilient Online Event AnlysisЭб IEEETransaction on Knowledge and Data Engineering,Vol.20,No.3,pp.289-299,2008