Рекомендательная система для ритейла: коллаборативная фильтрация с неявной обратной связью и повторяющимися транзакциями

Сарапулов Г. В.
Санкт-Петербургский государственный университет
Математико-механический факультет
g-eos@yandex.ru

4 апреля 2018 г.

Аннотация

В работе рассмотрено нескольких подходов к построению рекомендательной системы для продуктового ритейла: на основе вероятностной модели (model-based) и на основе сходства между объектами (neighborhood-based) и на основе матричных разложений (matrix factorization). Проведено сравнение этих подходов по точности предсказания покупок новых для покупателя товаров в будущем.

I. Введение

Рекомендательные системы предназначены для предсказания того, какие объекты могут быть интересны пользователю. Их сферы применения обширны (новостные и мультимедийные сервисы, поисковые системы, e-commerce и т. д.), и на фоне последних достижений в этой области и роста вычислительных мощностей и накопленных данных в последнее десятилетие наблюдается рост интереса бизнеса к таким системам.

Исследователи выделяют два основных типа рекомендательных систем. Системы, основанные на содержимом (content-based recommender systems) создаются таким образом, чтобы рекомендовать пользователю объекты, которые похожи на те, что он предпочитал в прошлом. Этот подход подразумевает оценку сходства между объектами на основе некоторых характеристик (например, для фильмов - жанр, режиссер, актеры). Часто бывает так, что описание объектов недоступно или является неполным, поэтому единственной доступной информацией является история транзакций между пользователями и объектами. Второй основной подход к построению рекомендательных систем - коллаборативная фильтрация (collaborative filtering) - основан на поиске закономерностей в истории транзакций, чтобы рекомендовать пользователям объекты, которые были релевантны для других пользователей с похожими предпочтениями.

Большая доля литературы посвящена обработке явной обратной связи (explicit feedback) при построении рекомендательных систем, во многом благодаря удобству работы с такими данными: задачу составления рекомендаций можно в этом случае свести к задаче классификации (на классы "нравится/не нравится") или к задаче регрессии (предсказывая рейтинг). На практике оценки объектов, как правило, недоступны, или их слишком мало. Это может быть связано с нежеланием пользователей выставлять оценки или с невозможностью получения подобной обратной связи. В этом более сложном случае имеется только неявная информация (implicit feedback) о предпочтениях пользователей: покупки, просмотры страниц сайтов, закономерности в поисковых запросах и т. п.

В работе рассмотрены несколько методов построения рекомендательной системы для продуктового ритейла, основанной на коллаборативной фильтрации в условиях неявной обратной связи, и проведена сравнительная оценка трех алгоритмов на чековых данных одной из торговых сетей. В силу специфики прикладной области сбор оценок товаров от пользователей на практике невозможен (неудобен для пользователей и дорого реализуем), а обработка персональной информации как правило требует наличия специального разрешения от пользователей, поэтому для выявления предпочтений

пользователей использовалась только анонимизированная история покупок. В разделе II изложена постановка задачи и описан принятый в работе способ разбиения данных для обучения и оценки качества рекомендательных систем. В разделе III рассмотрены три подхода к коллаборативной фильтрации:

- на основе ранжирования товаров по вероятности покупки в зависимости от наличия других товаров в корзине покупателя, для чего использовался наивный байесовский классификатор (model-based collaborative filtering);
- на основе сходства с товарами, которые покупатель приобретал ранее (item-to-item collaborative filtering);
- на основе матричных разложений (matrix factorization collaborative filtering).

В разделе IV описаны метрики качества рекомендательных систем и приведена сравнительная оценка моделей согласно этим метрикам.

II. Постановка задачи

Для формальной постановки задачи введем некоторые обозначения. Пусть U - множество покупателей, I - множество товаров, T - период, за который доступны транзакции, $R = \{r_{u,i,t}\}$ - множество транзакций (четверок покупатель-товар-дата-значение), S - множество возможных значений в транзакциях (например, $S = \mathbb{N}$, если транзакции представляю собой историю покупок, выраженную в количестве товаров или сумме трат, $S = \{0, 1\}$ - если используется информация только о фактах покупки/просмотра товара, или $S = \{1, 2, ..., 10\}$, если имеются явные оценки товаров по 10-ти бальной шкале). Для обозначения множества покупателей, которые приобретали товар i, будет использоваться U_i, аналогично, множество товаров, которые приобретал покупатель u, будет обозначаться I_u.

На основе предыдущих транзакций R необходимо для каждого покупателя $u \in U$ составить ранжированный список рекомендаций товаров $L(u) = [i_1, ..., i_k]$, которые могут быть ему интересны.

Поскольку рассматривается продуктовая розничная сеть, явные оценки товаров покупателями недоступны. Кроме того, покупки товаров могут повторяться и достаточно сильно зависят от сезонности. Таким образом, в отличие от большинства приложений рекомендательных систем (в новостных и мультимедийных сервисах, интернет-магазинах), здесь отсутствует необходимость рекомендовать пользователю исключительно новые для него товары (такие товары будем впоследствии обозначать как $I \setminus I_u$).

Для обучения и оценки рекомендательных систем множество транзакций R разбивается следующим образом. Множество покупателей U делится на обучающую U_{train} и тестовую U_{test} группы (в пропорции 4:1). Для каждого покупателя определяется дата t_u^*, определяющая конец исторического периода $T_u^{hist} = \{t \in T_u | t < t_u^*\}$, транзакции которого используются для предсказания транзакций будущего периода $T_u^{new} = T_u \setminus T_u^{hist}$. Например, такой датой может стать дата последней известной транзакции. Рекомендательные системы обучаются по транзакциям из множества R_{train}^{hist} предсказывать транзакции из множества R_{train}^{new}, качество полученной модели затем измеряется на тестовых выборках R_{test}^{hist} и R_{test}^{new}. Для оценки качества рекомендаций будем использовать метрики классификации precision и recall:

$$precision(L) = \frac{1}{|U|} \cdot \sum_{u \in U} \frac{|L(u) \cap T_u|}{|L(u)|} \tag{1}$$

$$recall(L) = \frac{1}{|U|} \cdot \sum_{u \in U} \frac{|L(u) \cap T_u|}{|T_u|} \tag{2}$$

III. Особенности неявной обратной связи

Yifan Hu et all [1] выделяют несколько отличительных особенностей данных о неявной обратной связи:

1. Отсутствие негативного отклика. По истории транзакций затруднительно выявить товары, которые неинтересны покупателю. Отсутствие покупок товара может говорить как о том, что покупателю неинтересен данный товар, так и о том, что покупатель не знает об этом товаре, или вообще приобретает его в других магазинах.

2. Неявный отклик зашумлен по своей природе. Покупка определенного товара еще не говорит о предпочтении покупателя: товар мог быть куплен в подарок или не понравиться покупателю.

3. Числовое значение отклика говорит не о степени предпочтения, а о степени уверенности. При неявном отклике численные значения отражают частоту взаимодействий, и большое значение само по себе не говорит о степени предпочтения: например, покупателю может нравится товар, который он покупает редко (например, из-за его дороговизны), и при этом он нейтрально относится к товару, который покупает постоянно. Однако числовое значение отклика все же является полезным: повторяющееся событие дает больше уверенности в том, что оно отражает реальные предпочтения покупателя, в то время как разовая покупка могла быть вызвана множеством других факторов.

4. Необходимость подбора подходящих метрик. При оценке систем с неявным откликом возникает сразу ряд нюансов: необходимо учитывать доступность товара, его взаимозаменяемость и взаимодополняемость с другими товарами, повторный отклик. В таких условиях применение стандартных метрик может быть неудовлетворительным.

IV. Методы коллаборативной фильтрации

i. Рекомендательная система на основе наивного байесовского классификатора

Пусть $\mathbf{x_u} = \{x_{u1}, ..., x_{uN}\}$ - вектор признаков покупателя u, построенный по истории транзакций R^{hist}, где $x_{ui} = 1$, если покупатель u покупал товар i, и $x_{ui} = 0$ в противном случае. Для каждого товара $i \in I$ обучается классификатор $f_i : \mathbf{X} \to [0, 1]$, оценивающий вероятность покупки товара i в зависимости от предыдущих покупок:

$$P(y_i = 1|\mathbf{x_u}) = \frac{P(y_i = 1) \cdot P(\mathbf{x_u}|y_i = 1)}{P(\mathbf{x_u})} \tag{3}$$

Наиболее релевантным товаром является тот, вероятность покупки которого максимальна:

$$i^* = \arg\max_{i \in I \setminus I_u} f_i(\mathbf{x_u}) \tag{4}$$

Для оценки вероятностей покупки товара i используем наивный байесовский классификатор:

$$P(y_i = 1|\mathbf{x_u}) = \frac{P(y_i = 1) \cdot \prod_{j=1}^{N} P(x_{uj}|y_i = 1)}{\prod_{j=1}^{N} P(x_{uj})} \tag{5}$$

Список рекомендаций для каждого покупателя ранжируется по убыванию вероятности покупки.

ii. Рекомендательная система на основе сходства товаров

Для определения сходства между двумя товарами i и j можно представить эти товары в виде векторов $\mathbf{x_i}$ и $\mathbf{x_j}$, где $x_{iu} = 1$, если покупатель u приобретал товар i, и $x_i = 0$, в противном случае. Компонентам этих векторов часто также присваивают вес, характеризующий степень предпочтения пользователем данного товара. Одним из вариантов таких весов могут быть TF-IDF [1] веса, позаимствованные из анализа текстов:

$$\textit{TF-IDF}(t_k, d_j) = \frac{f_{k,j}}{\max_z f_{z,j}} \cdot \log \frac{N}{n_k} \tag{6}$$

[1] Salton, G.: Automatic Text Processing. Addison-Wesley (1989)

где $f_{k,j}$ - частота встречания слова t_k в документе d_j, n_k - кол-в документов, где встречается слово t_k. Адаптируя этот подход для оценки весов товаров в корзине покупателя, примем за $f_{k,j}$ долю расходов на товар i_k в суммарных расходах покупателя u_j, за $\frac{n_k}{N}$ - долю товара i_k в обороте торговой сети. Веса дополнительно нормализуются:

$$w_{k,j} = \frac{TF\text{-}IDF(t_k, d_j)}{\sqrt{\sum_s^{|T|} TF\text{-}IDF(t_s, d_j)^2}} \tag{7}$$

Сходство товаров оценивается по косинусной мере:

$$sim(d_i, d_j) = \frac{\sum_k w_{k,i} \cdot w_{k,j}}{\sqrt{\sum_k w_{k,i}^2 \cdot \sum_k w_{k,j}^2}} \tag{8}$$

Для получения списка рекомендаций товары ранжируются по убыванию сходства с товарами, которые покупатель приобретал в прошлом.

iii. Рекомендательная система на основе матричных разложений

Альтернативный подход к коллаборативной фильтрации заключается в обнаружении латентных признаков, которые объясняют наблюдаемые данные. Наиболее распространены методы, основанные на сингулярном разложении матрицы наблюдений. В рамках модели каждому пользователю сопоставляется вектор $\mathbf{x_u} \in \mathbb{R}^f$, каждому объекту - вектор $\mathbf{y_i} \in \mathbb{R}^f$. Предсказание числовых значений откликов осуществляется через скалярное произведение $r_{ui} = \mathbf{x_u^\top y_i}$. Параметры такой модели в случае явных откликов оцениваются непосредственно по известным наблюдениям с регуляризацией:

$$\min_{x_*, y_*} \sum_{r_{ui} \in R} (r_{ui} - \mathbf{x_u^\top y_i})^2 + \lambda\Big(||x_u||^2 + ||y_i||^2\Big) \tag{9}$$

Для случая неявных откликов существуют несколько модификаций этого подхода. Согласно подходу, предложенному в Yifan Hu et all [1], в модель вводятся переменные, характеризующие предпочтения p_{ui} и степень уверенности c_{ui}:

$$\min_{x_*, y_*} \sum_{u,i} c_{ui}(p_{ui} - \mathbf{x_u^\top y_i})^2 + \lambda\Big(\sum_u ||\mathbf{x_u}||^2 + \sum_i ||\mathbf{y_i}||^2\Big)$$

$$p_{ui} = \begin{cases} 1, & \text{if } r_{ui} > 0 \\ 0, & \text{if } r_{ui} = 0 \end{cases} \tag{10}$$

$$c_{ui} = 1 + \alpha r_{ui}$$

Параметры α и λ зависят от входных данных и как правило оцениваются через кросс-валидацию. В работе хорошее качество показали значения $\alpha = 1$ и $\lambda = 0.01$.

После расчета латентных факторов пользователей и объектов в качестве рекомендаций для пользователя берутся объекты с наибольшими значениями предсказанного отклика $\hat{p}_{ui} = x_u^\top y_i$.

V. Результаты

Рекомендации, построенные с помощью подхода, основанного на вероятностной модели, оказались более точными, чем предсказания, основанные на сходстве товаров или на матричном разложении. В таблице 1 приведены значения точности (precision) для списков рекомендаций длины 1 и 3. Лучший результат naive bayes связан во многом с недостатками подхода, основанного на сходстве: в его рамках практически невозможно добиться появления в рекомендациях товаров, которые являются абсолютно новыми для покупателя, т. е. не имели аналогов в прошлых транзакциях.

Таблица 1: *Точность рекомендательных систем (precision at k)*

	Metric at k			
Model	Precision at 1	Precision at 3	Recall at 1	Recall at 3
Naive Bayes	0.40	0.35	0.02	0.05
Item2Item TF-IDF	0.27	0.22	0.02	0.05
Matrix Factorization via ALS	0.33	0.27	0.04	0.06

Выбранный в работе подход к оценке качества рекомендательных систем может быть улучшен. С одной стороны, имеет смысл сравнивать алгоритмы по метрикам ранжирования, т. к. в ряде приложений рекомендации представляют собой ранжированный список объектов, и более релевантные объекты должны в этом списке располагаться выше. С другой стороны, поскольку в рассматриваемой задаче пользователь покупает некоторые товары повторно, некоторые транзакции являются для него тривиальными, то есть он совершает их постоянно либо очень часто, и для более объективной оценки точности следует такие объекты исключить из тестовой выборки.

Список литературы

[Yifan Hu et all, 2008] Yifan Hu , Yehuda Koren , Chris Volinsky, Collaborative Filtering for Implicit Feedback Datasets, Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, p.263-272, December 15-19, 2008