УДК 004.78

Пятикоп Е.Е., к.т.н.

ИСПОЛЬЗОВАНИЕ СИНГУЛЯРНОГО РАЗЛОЖЕНИЯ МАТРИЦ В КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ

Приазовский государственный технический университет

Приведена классификация методов коллаборативной фильтрации. Описаны математические основы метода понижения размерности измерений на основе сингулярного разложения матриц (SVD). Представлен подход нормализации данных с использованием базовых прогнозов. Показаны результаты экспериментов реализации метода

Введение

В настоящее время всемирная паутина – глобальное хранилище информации, объем которой постоянно увеличивается. Поэтому часто пользователи сталкиваются со сложной задачей поиска информации, которая бы соответствовала их предпочтениям. Для решения этой задачи используют рекомендательные системы, которые автоматически предоставляют рекомендации пользователям на основании уже совершенных действий (покупок, выставленных рейтингов, посещений и т.д.) и результатах обратной связи (заказы в магазинах, переход по ссылкам и т.п.). В первую очередь такие системы востребованы в электронной коммерции, но их использование может быть расширено на справочные центры, поиск по программному обеспечению, научным статьям и т.п. Это приведет к значительной экономии времени потребителя на поиски необходимого объекта.

Одним из подходов разработки рекомендательных систем является использование методов коллаборативной фильтрации (КФ). Коллаборативная фильтрация – класс методов построения рекомендаций (прогнозов) на основе известных предпочтений (оценок) группы пользователей. Например, ресурс Либрусек представляет более 290 тыс. книг и каждый месяц более 3000 обновлений [1]. Пользователи оценивают прочитанные книги. На основе этих оценок другие читатели могут определить, понравится ли им эта книга. Для оптимизации данного процесса возможно использование рекомендательных систем на основе КФ.

Основная идея алгоритмов КФ заключается в предложении новых элементов для конкретного пользователя на основе предыдущих предпочтений пользователя или мнений других единомышленников пользователя.

На сегодняшний день исследователи разработали целый ряд алгоритмов КФ [2-3], которые можно разделить на следующие основные категории:

1. Методы, основанные на анализе имеющихся оценок, – анамнестические [1] методы *(Memory-based).* Эти алгоритмы опираются на статистические методы, чтобы найти группу пользователей близких к целевому пользователю. Этот подход еще называют метод ближайших соседей: использование предшествующих оценок, сделанных клиентом, и анализ оценок других пользователей, которые имеют подобные предпочтения. Тогда рекомендации (прогноз) для целевого пользователя формируются на основании вычисления некой меры похожести по всем накопленным данным.

2. Методы, основанные на анализе модели данных, – модельные методы *(Model-based).* В этом случае сначала по совокупности оценок формируется описательная модель предпочтений пользователей, товаров и взаимосвязи между ними, а затем вырабатываются рекомендации на основании полученной модели. Процесс формирования рекомендаций разбивается на два этапа: обучение моде-

[1] АНАМНЕСТИЧЕСКИЙ, АНАМНЕЗ [нэ], -а, м. (спец.). Совокупность медицинских сведений, получаемых путем опроса обследуемого и знающих его лиц.

ли в отложенном режиме и достаточно простое вычисление рекомендаций на основе существующей модели в реальном времени. Эти алгоритмы могут быть основаны на вероятностном подходе [3], кластерном анализе [4], анализе скрытых факторов [5].

3. Методы, основанные на объединении предыдущих алгоритмов, – гибридные методы.

Эти подходы в свою очередь могут быть еще разбиты далее на группы методов. Так, методы первой группы, основанные на соседстве (близости), разделяются на методы анализа сходства пользователей (*User-based*) или сходства элементов (*Item-based*). Целью обоих направлений является выделение схожих объектов в группы на основе матрицы оценок [1-3]. В первом случае определяется сходство пользователей: найти других пользователей, чьи прошлые оценки поведения похожи на те, что и у текущего пользователя, и использовать их оценки элементов для прогнозирования предпочтения текущего пользователя. Второй подход основан на сходстве элементов [6, 7]. Если два элемента, как правило, имеют одинаковые оценки пользователей, то они похожи, и пользователи должны иметь аналогичные предпочтения для подобных элементов.

Коллаборативная фильтрация на основе сходства пользователей (*User-based*) имеет высокую точность. Однако недостатком является ресурсоемкость (требование к памяти) и сложность (количество вычислений, требуемое для получения рекомендаций). К тому же вычисление степени близости может производиться только в реальном времени, так как данные о текущей транзакции становятся доступными только в момент выработки рекомендаций. Поэтому данный метод может применяться только к относительно небольшим базам данных.

В алгоритме на основе сходства элементов (*Item-based*) степень близости анализируемого элемента ко всем остальным может быть вычислена в отложенном режиме. Поэтому этот алгоритм оказывается более эффективным с точки зрения времени формирования рекомендаций благодаря возможности проведения отложенной предобработки данных.

Несмотря на легкость понимания и реализации описанных выше методов, у них есть ряд недостатков. Во-первых, возникают трудности при прогнозе предпочтений для новых пользователей или при появлении новых элементов, т.к. для них еще нет оценок. Во-вторых, для расчета предпочтений необходимо хранить всю матрицу данных. В-третьих, ограничивается возможность методов при обработке больших объемов данных, т.к. выполнение большого количества операций для вычисления степени похожести затрудняет выдачу рекомендаций в реальном времени. Большой объем матрицы предпочтений затрагивает также проблему избыточности данных. Как правило, пользователи и элементы делятся на группы с аналогичными профилями предпочтений. Например, многие научно-фантастические фильмы будут нравиться в аналогичной степени тем же наборам пользователей.

Вследствие этого, возникает задача в понижении размерности матрицы оценок. Такие задачи решают методы второй группы модельные методы *(Model-based)*. Эти методы направлены на поиск закономерностей на основе обучающих данных. Такой подход является комплексным и даёт более точные прогнозы, так как помогает раскрыть скрытые факторы, объясняющие наблюдаемые оценки.

В этом случае возможен вариант объединения пользователей (элементов) в кластеры (профили) с помощью некоторого индекса сходства. Элементы и оценки, выставленные пользователями из одного кластера, будут использоваться для вычисления рекомендаций. Кластерные модели лучше масштабируются, т.к. сверяют профиль пользователя с относительно небольшим количеством сегментов, а не с целой пользовательской базой. Сложный и емкий кластерный подсчет

ведется в оффлайн режиме. Эта задача может быть выполнена на основе разных математических подходов [4, 5], автором выбран путь сокращения размерности с помощью сингулярного разложения матриц (*Singular value decomposition или SVD*) [8, 9].

Постановка задачи

Информационная область для систем КФ состоит из пользователей, которые выразили предпочтения для различных предметов. Предпочтение (оценка) часто представляется в виде триплета (пользователь, предмет, оценка). Эти оценки могут принимать различные формы, в зависимости от рассматриваемой системы. Некоторые системы используют вещественную или целочисленную оценочную шкалу, такую как 0-5 звезд, другие используют бинарные или тройные меры. Множество всех триплетов оценок формирует разреженную матрицу, называемую матрицей оценок. Пары (Пользователь, предмет), в которых пользователи не отдали предпочтение предмету, являются неизвестными значениями этой матрицы (Табл. 1).

Таблица 1. Матрица оценок

	Элемент 1	**Элемент 2**	**Элемент 3**
Пользователь 1	3	?	2
Пользователь 2	?	4	3
Пользователь 3	5	4	?

При использовании системы КФ необходимо решить две задачи: 1) спрогнозировать оценку или предпочтение, которое пользователь отдаст предмету; целью прогноза является заполнение в матрице оценок недостающих значений; 2) выдача рекомендации, т.е. формирование ранжированного списка N элементов для данного пользователя.

Определим математические обозначения для привязки различных элементов модели рекомендательных систем. Генеральная совокупность состоит из набора пользователей U и набора элементов I. I_u - множество элементов, оцененных пользователем u. U_i - множество пользователей, которые оценили элемент i. $r_{u,i}$ - оценка пользователя u для элемента i. r_u - вектор всех оценок пользователя u. r_i - вектор всех оценок элемента i. $\bar{r}_u$ и $\bar{r}_i$ - средние значения оценок пользователя u и элемента i соответственно. Рекомендательный прогноз – $r_{u,i}$.

Сингулярное разложение матриц (SVD)

Суть метода в разложении матрицы $A \in M(n,m)$ с рангом $d = rank(M) \le \min(n,m)$ в произведение матриц меньшего ранга:

$$A = UDV^T, \qquad (1)$$

где матрицы $U \in M(n,d)$ и $V \in M(m,d)$ состоят из ортонормальных столбцов, являющихся собственными векторами при ненулевых собственных значениях матриц AA^T и A^TA соответственно и $U^TU = V^TV = I$, а $D \in M(d,d)$

$D = \begin{pmatrix} a_1 & 0 & 0 \\ 0 & . & 0 \\ 0 & 0 & a_d \end{pmatrix}$ - диагональная матрица с положительными диагональными элементами $a_1 \ge a_2 \ge \mathbf{K} \ge a_d > 0$, отсортированными в порядке убывания.

Диагональные элементы матрицы D $a_1, a_2, \mathbf{K}, a_d$ представляют собой собственные значения, соответствующие ненулевым собственным векторам AA^T и A^TA (столбцам U и V). Столбцы матрицы U представляют собой, ортонормальный базис пространства столбцов матрицы A, а столбцы матрицы V – ортонормальный базис пространства строк матрицы A. Важным свойством SVD-разложения является тот факт, что если для $k<d$ преобразовать матрицу D в матрицу

$D_k = \begin{pmatrix} a_1 & 0 & 0 \\ 0 & . & 0 \\ 0 & 0 & a_k \end{pmatrix} \in M(k,k)$, состоящую только из k наибольших диагональных элементов, а также оставить в матрице U и V только k первых столбцов, т.е. преобразовать их в $U_k \in M(n,k)$ и $V_k \in M(m,k)$, то матрица

$$A_k = U_k D_k V_k^T \qquad (2)$$

будет являться лучшей аппроксимации матрицы A относительно нормы Фробениуса среди всех матриц с рангом k, т.е. $\|A - A_k\| \le \|A - A'\| \quad \forall A' \in M(n,m)$, $rank(A') = k$.

Это усечение одновременно достигает двух целей. Во-первых, оно уменьшает размерность векторного пространства, снижает требования хранения и вычислительные требования к модели. Во-вторых, отбрасывая малые сингулярные числа, малые искажения в результате шума в данных удаляются, оставляя только самые сильные эффекты и тенденции в этой модели. Снижение воздействия шума улучшает способность предоставлять высококачественные рекомендации.

Применительно к задаче коллаборативной фильтрации сингулярные разложения можно использовать в таком порядке:

1. Построить разложение (1);

2. Зафиксировать некоторое число скрытых факторов k, которое, так или иначе, описывает каждый элемент и предпочтения каждого пользователя относительно этих факторов. При выборе учесть $k << rank(A)$. Можно подбирать k, исходя из размера сингулярных значений матрицы, т.е. тех самых диагональных элементов матрицы D: желательно отбрасывать как можно больше, но при этом как можно более маленьких таких элементов;

3. Получить лучшую k-ранговую аппроксимацию матрицы А в форме (2).

Когда преобразование (2) завершено, пользователи и элементы могут быть представлены в виде точек в k-мерном пространстве. Представляем каждого пользователя вектором из k факторов r_u и каждый продукт вектором из k факторов r_i, чтобы предсказать рейтинг пользователя u товару i, вычисляем их скалярное произведение:

$$r_{u,i} = r_i \cdot r_u = r_i^T \cdot r_u \qquad (3)$$

Можно сказать, что вектор факторов пользователя показывает, насколько пользователю нравится или не нравится тот или иной фактор, а вектор факторов продукта показывает, насколько тот или иной фактор в продукте выражен.

Нормализация данных

Как правило, подавляющее большинство оценок неизвестно, и разреженность матрицы оценок достаточно высока. С другой стороны, данные, которые уже имеются в матрице, достаточно субъективны. Некоторые пользователи – оптимисты, и их оценки всегда высоки (среднее 4 из 5), другие пользователи – циники, их оценки всегда занижены (среднее 2,5 из 5). Кроме этого, всегда есть элементы, которые нравятся всем.

В целях борьбы с подгонкой разреженных данных оценок, проводится регуляризация моделей таким образом, чтобы сократить вероятность появления случайных связей между оценками, которые не отражают действительность. Регуляризация контролируется константами, которые обозначаются как λ_1, λ_2, ... Точные значения этих констант определяются перекрестной проверкой. По мере их роста, регуляризация становится все тяжелее.

Для того чтобы оптимизировать производительность выдачи рекомендаций, важно нормализовать оценки до вычисления матрицы подобия. Это может быть достигнуто путем вычисления базового прогноза (предикатора), в котором инкапсулируют отклонение пользователя и элемента. Пары пользователь-элемент (u,i), для которых оценки $r_{u,i}$ известны, составляют

множество K. Базовый прогноз для неизвестной оценки $r_{u,i}$ обозначается $b_{u,i}$ и определяется формулой:

$$b_{u,i} = m + b_u + b_i \qquad (4)$$

где μ - общая средняя оценка; b_u и b_i - параметры, которые показывают наблюдаемое отклонение пользователя u и элемента i соответственно от среднего значения.

Теперь выразим предпочтения (3) с учетом базового прогноза (4).

$$\hat{r}_{u,i} = m + b_u + b + r_i^T \cdot r_u \qquad (5)$$

Так как все параметры (3) взаимосвязаны, то рассчитывать их необходимо вместе, решив задачу наименьших квадратов.

$$\sum_{(i,u)\in D} (\hat{r}_{u,i} - m - b_u - b - r_i^T \cdot r_u)^2 \qquad (6)$$

Для минимизации данного выражения используется градиентный спуск: берем частные производные по каждому аргументу и двигаемся в сторону, обратную направлению этих частных производных.

$$\min \sum_{u,i\in K} (r_{u,i} - m - b_u - b_i - p_u^T \cdot q_i)^2 +$$

$$+ l_3(\|p_u\|^2 + \|q_i\|^2 + b_u^2 + b_i^2)$$

$$q_i \leftarrow q_i + g_2(e_{u,i} \cdot p_u - l_2 \cdot q_i)$$

$$p_u \leftarrow p_u + g_2(e_{u,i} \cdot q_i - l_2 \cdot p_u)$$

$$b_u \leftarrow b_u + g_1 \cdot (e_{i,j} - l_1 \cdot b_u)$$

$$b_i \leftarrow b_i + g_1 \cdot (e_{i,j} - l_1 \cdot b_i)$$

$$e_{i,j} = r_{i,j} - \hat{r}_{i,j}$$

где γ_1, γ_2, λ_1, λ_2, λ_3 – константы регуляризации.

На основе представленных математических выкладок были проведены исследования по использованию описанного метода. Для оценки эффективности использовалась среднеквадратичное отклонение.

$$RMSE = \sqrt{\frac{1}{n}\sum_{u,i} \left(\hat{r}_{u,i} - r_{u,i}\right)^2} \qquad (7)$$

где $r_{u,i}$ – известная оценка пользователя u для элемента i, $\hat{r}_{u,i}$ – спрогнозированная оценка.

Результаты исследований

В качестве исходных данных использовались таблицы базы данных с оценками пользователей объемом – 20000 оценок, с данными о пользователях – 200 строк и данными о книгах – 1500 строк.

Первым из этапов выдачи рекомендаций является обучение модели расчетов. Качество обучения определяется погрешностью обучения и выражается среднеквадратичным отклонением прогнозируемых оценок. Для качественного обучения модели необходимо подобрать параметры обучения таким образом, чтобы погрешность была наименьшей в соответствии с заданной точностью.

Исследуемый параметр модели – количество скрытых факторов k. Другие параметры исследованы [9] и приняты: коэффициент скорости обучения базового отклонения оценок $g_1 = 0.007$, коэффициент скорости обучения факторов $g_2 = 0.007$, коэффициент регуляризации базового отклонения оценок $l_1 = 0.005$, коэффициент регуляризации факторов $l_2 = 0.015$. Приняв за основу эти данные, определи оптимальное количество факторов, при точности ε=0,00001. Результат показан в таблице 2.

Таблица 2. Изменение количества скрытых факторов k

№ п/п	Количество факторов k	RMSE при обучении	RMSE при тестировании
1	100	0.10303	0.95778
2	200	0.09727	0.94213
3	300	0.09598	0.93274
4	400	0.09479	0.92988
5	500	0.09204	0.92742
6	600	0.08930	0.92536
7	700	0.08655	0.92380
8	800	0.08381	0.92164
9	900	0.08106	0.91978
10	1000	0.07831	0.91862
11	1100	0.07557	0.91776
12	1200	0.07282	0.91640
13	1300	0.07008	0.91564
14	1400	0.06733	0.91358
15	1500	0.05236	0.91243

Согласно полученным результатам можно сделать вывод о том, что с

увеличением числа факторов, погрешность прогнозов уменьшается. Для исходных данных оптимальным количеством факторов является k =1500. Однако с увеличением числа факторов возрастают вычислительные затраты и время обучения.

В работе [7] автором был рассмотрен метод коллаборативной фильтрации на основе сходства элементов (Item-based). Дальнейшие исследования были направлены на сравнение результатов этого метода с методом, рассмотренным в настоящей статье. Немаловажным фактором является время обучения модели. Результаты этих исследований показаны в таблице 3.

Таблица 3. Сравнение по времени

№ п/п	Объем обучающих данных	Время обучения (чч:мм:сс)	
		SVD	Item-based
1	2500	0:09:17	0:23:33
2	5000	0:15:29	0:45:12
3	7500	0:21:42	1:03:25
4	10000	0:27:54	1:32:16
5	12500	0:34:07	1:54:47
6	15000	0:41:15	2:19:14
7	17500	0:58:52	2:57:26
8	20000	1:12:24	3:29:21

Результаты подтвердили, что уменьшение размерности матрицы оценок значительно сокращает время обучения модели. Исследования по изменению качества показали, что RMSE с использованием SVD уменьшилась на 2%. Учитывая, что время при этом уменьшилось почти в 3 раза, то использование сингулярного разложения применительно к рассматриваемой задаче действительно эффективно.

Список литературы

1. Либрусек- Статистика. Режим доступа: http://lib.rus.ec/stat
2. Xiaoyuan Su and Taghi M. Khoshgoftaar A Survey of Collaborative Filtering Techniques A Survey of Collaborative Filtering Techniques // Hindawi Publishing Corporation, Advances in Artificial Intelligence archive, USA: 2009. – С. 1-19.
3. Гомзин А. Г., Коршунов А. В. Системы рекомендаций: обзор современных подходов // Труды ИСП РАН. 2012. №. С.401-418.
4. Савчук Т.О., Сакалюк А.В. Застосування кластерного аналізу для колаборативної фільтрації / Т.О. Савчук, // Вісник Хмельницького національного університету. –2011 – №1– С. 186-192
5. Лексин В.А., Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов // Математические методы распознавания образов-13. – М. МАКС Пресс, 2007. – С. 488-491
6. Sarwar B. M. Item-based collaborative filtering recommendation algorithms / B. M. Sarwar, G. Karypis, J. A. Konstan // Proceedings of ACM WWW '01, pp. 285–295, ACM, 2001.
7. Пятикоп Е.Е. Исследование метода коллаборативной фильтрации на основе сходства элементов // Наукові праці Донецького національного технічного університету серія: "Інформатика, кібернетика та обчислювальна техніка". – 2013. – №2. – С. 109-114
8. Vozalis1 M. G., Margaritis K. G. Applying SVD on Generalized Item-based Filtering //International Journal of Computer Science & Applications Vol. 3 Issue 3, pp 27- 51
9. Koren Y., Ave P., Park F. Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model. In: KDD '08 Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining Режим доступа: http://public.research.att.com/~volinsky/netflix/kdd08koren.pdf

Статью представлено в редакцию 9.12.2013