Оценка эффективности рекомендательных систем

© С.А.Амелькин
Институт программных систем имени А.К.Айламазяна РАН,
Переславль-Залесский
sam@sam.botik.ru

Аннотация

Рекомендательные системы позволяют на основании выставленных пользователями оценок объектам прогнозировать значения оценок и осуществлять автоматический выбор объектов, которые могут заинтересовать пользователя. В последнее время разработано большое количество различных рекомендательных систем, однако формальной постановки задач, решаемых такими системами, нет. В работе рассмотрены три задачи, которые могут быть решены методами коллаборативной фильтрации и показано, как постановка задач зависит от спектра шкалы оценок и понятия близости оценок, выставленных разными пользователями.

1 Введение

В настоящее время с развитием интернет-сервисов – магазинов, библиотек, средств массовой информации – большое внимание уделяется рекомендательным системам. Однако, формальной задачи, которая решается с помощью метода, положенного в основу рекомендательных систем: коллаборативного подхода, нет. Вместо постановки задачи предлагается набор критериев, не являющихся взаимозаменяемыми, а, значит, соответствующих различным задачам, которые решаются с использованием одного и того же метода, называемого коллаборативным подходом [1]. Этот метод основан на следующих неявных предположениях:

- Чем больше совпадений между оценками двух пользователей рекомендательной системы, тем больше вероятность того, что и остальные оценки у этих пользователей совпадают.
- Можно выбрать операцию усреднения, позволяющую по оценкам объекта, сделанным множеством пользователей, восстановить оценку этого же объекта для конкретного пользователя.
- Оценки объектов пользователями не меняются во времени, транзитивны и не подвержены влиянию случайных помех.

Поскольку объекты оцениваются в порядковой шкале, вид операции усреднения зависит от спектра шкалы, а также от способа определения совпадений оценок пользователей. Действительно, совпадения оценок могут быть рассчитаны по значению оценки, попаданию в определенные семантические страты, совпадению порядков, построенных по множеству оценок и пр. Таким образом, использование одних и тех же предположений не приводит к однозначности поставленной задачи. А значит, рекомендации, полученные решением различных задач, не могут быть оценены одними и теми же показателями эффективности. Более того, в рамках некоторого показателя эффективности результаты работы различных рекомендательных систем могут оказаться несравнимыми. Среди разработчиков рекомендательных систем существует понимание этого факта, поэтому для измерения эффективности работы рекомендательных систем введен [2] ряд различных показателей (Табл. 1). Аббревиатуры показателей эффективности будут расшифрованы ниже. Такое разнообразие показателей показывает общее состояние разработок в этой области, когда на основе немногочисленных признанных достижений работают параллельно много независимых групп, создающих свою терминологию, методы верификации и проверки полученных результатов.

Цель работы – дать формальную постановку для задач, решаемых с использованием коллаборативного подхода, рассмотреть возможные критерии, экстремум которых соответствует максимальной эффективности рекомендательных систем, исследовать возможности решения многокритериальных задач в этой области.

Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации.

2 Формальные постановки задач

Пусть задана шкала оценок. Если спектр оценок содержит достаточно много возможных значений (учитывая, что оценка всегда дискретна), то невозможно выделить семантически каждую оценку.

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

Таблица 1 *Критерии проверки гипотез*

precision positive prediction value	точность (точность, усредненная по пользователям системы имеет обозначение map.)	$prc = \frac{tp}{tp + fp}$
recall (recall rate), rtu positive rate, sensitivity, hitrate coverage	полнота	$rcl = \frac{tp}{tp + fn}$
negative prediction value	точность отрицательного прогноза	$npv = \frac{tn}{tn + fn}$
specificity	специфичность	$spc = \frac{tn}{tp + fn}$
accuracy	аккуратность	$acc = \frac{tp + tn}{n}$ 1 - β-error =α-error +acc; $acc = \alpha\, rcl + (1\text{-}\alpha) spc$
F-measure (F score)	F мера (мера Ван Ризбергена)	$F\beta = \frac{(1+\beta^2) prc \cdot rcl}{\beta^2 prc + rcl} = \left(\frac{\alpha}{prc} + \frac{(1-\alpha)}{rcl}\right)^{-1}$ $\beta^2 = \frac{1-\alpha}{\alpha}$
false discovery rate		$fdr = \frac{fp}{tp + fp}$
false positive rare fall-out		$fpr = \frac{fp}{tn+fn}$; $\left(\frac{1}{fpr} + \frac{1}{fdr}\right)^{-1} = \alpha\text{-}error$

Поэтому оценки объединяются в семантические страты. По результатам оценок можно построить отношение порядка на множестве оцененных объектов. Таким образом, есть три возможности указать на совпадение оценок двух пользователей: по совпадению или незначительной разности значений оценки, по совпадению страт, в которые попадают эти оценки, по совпадению упорядоченных множеств объектов. В соответствии с каждой из этих возможностей, можно выделить три задачи рекомендательной системы.

2.1 Задача прогноза оценки объекта пользователем

Обозначим:

r_i – результат работы рекомендательной системы: оценка в принятой шкале, сгенерированная системой в ходе i-го обращения к ней (вне зависимости от пользователя и объекта).

n – число обращений к системе (объем выборки).

v_i – соответствующая i-ому обращению к системе собственная оценка пользователя, данная им объекту; v_i не известна алгоритму и может быть использована в ходе испытаний за счет выделения контрольного множества оценок.

Предполагается, что величины v_i не зависят от r_i, они постоянны во времени (эти предположения могут быть сняты при уточнении модели пользователя).

<1> Задачей рекомендательной системы является расчет значений r_i, максимально близких к величинам v_i, при заданном множестве пар (пользователь, объект), доля которых известны v_i.

Для любого фиксированного n результат работы рекомендательной системы: вектор $R=(r_1, r_2, \ldots, r_n)$ представляет собой точку в пространстве M^n, где M – множество (спектр) оценок, допустимых в используемой шкале. Этому же пространству принадлежит точка $V=(v_1, v_2, \ldots, v_n)$. Формальная постановка задачи <1>

$$d(R,V) \to \min, \tag{1}$$

где d – метрика, выбранная на пространстве M^n. Если на пространстве M^n выбрана норма $\|V\|$, например, из класса lp, то соответствующее этой норме расстояние

$$d(R,V) = \|R - V\| = \left(\sum_{i=1}^{n} |r_i - v_i|^p\right)^{\frac{1}{p}} \tag{2}$$

может служить критерием эффективности для задачи (1). Нормируем расстояние $d(R,V)$, чтобы результат не зависел от объема выборки n, $d'(R,V) = \frac{1}{n^{\frac{1}{p}}} d(R,V)$, получим:

$$p = 1 \Rightarrow d'(R,V) = mae = \frac{1}{n}\sum_{i=1}^{n} |r_i - v_i|$$

$$p = 2 \Rightarrow d'(R,V) = rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (r_i - v_i)^2}$$

Отметим, что в такой постановке не имеет значения знак разности между v_i и r_i , а также

величина оценки. Выбор параметра *p* осуществляется из соображений значимости больших отклонений по сравнению с малыми.

2.2 Задача определения страты, в которую попадает оценка

Разделим принятую в данной рекомендательной системе шкалу на две страты: *P* – положительные оценки и *N* – отрицательные оценки.

<2> Тогда работа рекомендательной системы представляет собой проверку гипотезы H_0:$v_i \in P$.

Так же, как и для любого процесса проверки гипотезы, результат работы можно свести (табл. 2)

Таблица 2 *Возможные результаты прогноза*

	$v_i \in P$	$v_i \in N$
$r_i \in P$	положительный прогноз верен	ошибка 1 рода (α - ошибка)
$r_i \in N$	ошибка 2 рода (β - ошибка)	отрицательный прогноз верен

Обозначим:

tp – число опытов, в которых положительный прогноз оказался верен,

tn – число опытов в которых отрицательный прогноз оказался верен,

fp – число опытов, в которых положительный прогноз оказался ошибочным (число опытов в котороых наблюдалась ошибка первого рода),

fn – число опытов, в которых отрицательный прогноз оказался ошибочным (число опытов, в которых наблюдалась ошибка второго рода).

Сравнивая величины *tp*, *tn*, *fp*, *fn* (*tp*+*tn*+*fp*+*fn*=*n*), можно получить различные критерии для формализации задачи <2> (табл. 3).

Отметим, что эти критерии могут быть противоречивыми: уменьшение уровня значимости (*α-error*) возможно только за счет снижения мощности критерия проверки гипотезы (1 – *β-error*). Если в качестве критериев деятельности рекомендательной системы выбраны несколько таких противоречивых критериев, то настройки системы могут обеспечить увеличение эффективности по одному из критериев только за счет ухудшения эффективности по другому, таким образом составляя множество Парето. Такое множество для рекомендательных систем строится, как правило либо для показателей prc(rcl), либо для rcl(fpr).

Для свертки критериев могут использоваться различные функции двух или более критериев, такие, как взвешенные суммы, усредненные оценки типа fall-out или, например, коэффициент корреляции Мэтью:

$$mcc = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp+fn)(fp+tn)(tp+fp)(tn+fn)}}$$

Таблица 3 *Критерии проверки гипотез*

precision positive prediction value точность[1]	$prc = \frac{tp}{tp+fp}$
recall (recall rate), rtu positive rate, sensitivity, hitrate coverage полнота	$rcl = \frac{tp}{tp+fn}$
negative prediction value точность отрицательного прогноза	$npv = \frac{tn}{tn+fn}$
specificity специфичность	$spc = \frac{tn}{tp+fn}$
Accuracy аккуратность	$acc = \frac{tp+tn}{n}$ 1 - β-error =α-error +*acc* *acc* = α *rcl* + (1-α)*spc*
F-measure (F score) F мера (мера Ван Ризбергена)	$F\beta = \frac{(1+\beta^2)prc \cdot rcl}{\beta^2 prc + rcl} = \left(\frac{\alpha}{prc} + \frac{(1-\alpha)}{rcl}\right)^{-1}$ $\beta^2 = \frac{1-\alpha}{\alpha}$
false discovery rate	$fdr = \frac{fp}{tp+fp}$
false positive rare fall-out	$fpr = \frac{fp}{tn+fn}$ $\left(\frac{1}{fpr} + \frac{1}{fdr}\right)^{-1} = \alpha$-error

Существует отличие между задачами <1> и <2>. Оно заключается в том, что в первой задаче пространство оценок считается однородным, тогда как во второй оно разграничено стратами. При проверке гипотез две оценки, расположенные по обе стороны от границы страт сколь угодно близко друг к другу, оказываются значительно «дальше», чем оценки разность которых больше, находящиеся в одной страте.

Использование рассмотренных критериев возможно и при числе страт, большем двух. В этом случае следует выделить проверяемую гипотезу о попадании прогнозной оценки в заданную страту. С другой стороны, при количестве страт, большем двух, можно преобразовать шкалу оценок так, чтобы каждой страте соответствовала одна оценка. Тогда можно формализовать задачу рекомендации объекта в виде задачи <1>.

2.3 Задача выбора последовательности лучших объектов

Оценки всегда формируют линейно упорядоченное множество: для любых *i*, *j* выполняется либо

[1] Точность, усредненная по пользователям системы имеет обозначение *map*.

$r_i \geq r_j$ либо $r_j \geq r_i$. Аналогичное утверждение верно и для множества оценок v_i.

<3> Задача рекомендательной системы – найти такие оценки r_i, чтобы для максимального числа пар (i, j) (i=1, …, n; j=1, … ,n; $i \neq j$) *выполнялось условие:*

$$\begin{cases} r_i > r_j & \Leftrightarrow \quad v_i > v_j \\ r_i = r_j & \Leftrightarrow \quad v_i = v_j \\ r_i < r_j & \Leftrightarrow \quad v_i < v_j. \end{cases}$$

Критерием эффективности для задачи <3> может служить отношение числа правильно построенных порядков. В частности, к таким критериям относятся

расстояние Хэмминга (fraction of concordant pairs, dH): для всех пар (i, j) записываются двоичные коды

$$C_{ij}^{v} = \begin{cases} 1, & \text{если} \quad v_i \geq v_j; \\ 0, & \text{если} \quad v_i < v_j; \end{cases} \quad (v \in \{r, v\});$$

$$dH = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(C_{ij}^{r} \leftrightarrow C_{ij}^{v}\right).$$

Корреляция Пирсона

$$r = \frac{n \sum r_i v_i - \sum r_i \sum v_i}{\sqrt{n \sum r_i^2 - (\sum r_i)^2} \sqrt{n \sum v_i^2 - (\sum v_i)^2}}; \qquad (3)$$

При расчете корреляции Пирсона используются центральные моменты, представляющие средние значения (по множеству опытов) разностей величин и их средних значений.

Косинус угла между векторами R и V:

$$\cos \varphi = \frac{\sum r_i v_i}{\sqrt{\sum r_i^2} \sqrt{\sum v_i^2}}; \qquad (4)$$

Выражение для cos φ аналогично корреляции r, только для вычисления косинуса используются не центральные, а начальные моменты.

Непараметрические коэффициенты корреляции (Кендалла, Спирмена, Фехнера и пр.)

Нормированный показатель дисконтированной накопленной выгоды (я не нашел точного русского соответствия термину normalized discounted cumulative gain) — ndcg, — рассчитываемый, как отношение

$$ndcg = \frac{dcg(r)}{dcg(v)}, \text{где } dcg(v) = v_1 + \sum_{i=1}^{n} \frac{v_i}{\log_2 i}.$$

Решение задачи <3> позволяет выделить последовательность объектов, ожидаемая оценка для которых является наивысшей из всего множества оцениваемых объектов. Однако, такой прогноз не гарантирует значение наивысших оценок или страту, в которую эти оценки попадут. Поэтому целесообразно рассматривать задачу прогноза оценки, как многокритериальную задачу, в которой критериями являются критерии задач <1> - <3>.

Литература

[1] L.H. Ungar and D.P. Foster. Clustering Methods for Collaborative Filtering, AAAI Workshop on Recommendation Systems, 1998.

[2] 5th ACM International Conference on Recommender Systems. Bilbao, October 2011.

Evaluation of Recommender Systems Efficiency

Sergey Amelkin

Development of internet services leads to investigation of forecast algorithms. One of them is collaboration filtering. It allows a user to find objects to be the most interesting for him/her. Unfortunately, there is no formal statement of problems solved by the collaboration filtering algorithms. Three problems are considered in the paper, differences between the statements are discussed. It is shown that a spectrum of the valuation scale and definition of notion of proximity of valuations provided by different users determine the type of problem to solve.