Leonardo Filgueira

Sistemas de recomendação usando o software R

Niterói - RJ, Brasil

Universidade Federal Fluminense

Leonardo Filgueira

Sistemas de recomendação usando o software R

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Prof. Luciane Ferreira Alcoforado

Niterói - RJ, Brasil

Universidade Federal Fluminense

Leonardo Filgueira

Sistemas de recomendação usando o software R

Monografia de Projeto Final de Graduação sob o título *"Sistemas de recomendação usando o software R"*, defendida por Leonardo Filgueira e aprovada em , na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Luciane Ferreira Alcoforado
Departamento de Estatística – UFF

Prof. Dr. Steven Dutt Ross
UNIRIO

Prof. Dr. Nome do 2o membro da banca
Instituicao do 2o membro da banca

Niterói,

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A partir do aumento de informação disponível com a popularização da Internet e com a possibilidade de armazenar essas informações, surge o desafio de lidar com este grande conjunto de dados[1]. Este aumento de informações desafia o site, como lojas on-line, que recebe todas as informações dos usuários que visitam o endereço, mas também pode se tornar um problema para o usuário que, diante da grande quantidade de produtos disponíveis para compra, pode levar muito tempo para achar o produto desejado[2].

Sistemas de recomendação são técnicas de *machine learning* que filtram um grande conjunto de dados, tendo como base informações dos usuários[3]. A partir dessas técnicas são previstas as notas que os usuários dariam a determinados itens, que podem ser dos mais variados tipos, e, para um indivíduo, recomenda-se o(s) item(ns) que obtiveram uma nota prevista maior[4]. Os sistemas de recomendação têm como objetivo recomendar itens que interessariam aos usuários[5], beneficiando o usuário e a loja, pois eles aumentam o desempenho da loja, fazendo-a vender uma quantidade maior de produtos, e também facilitam a procura do usuário fazendo-o achar produto(s) desejados em um menor tempo[1].

O primeiro sistema de recomendação foi criado na década de 90 e tinha como nome "filtragem colaborativa", pois o sistema funcionava com base na colaboração entre os grupos de pessoas interessados. Contudo, o termo "sistemas de recomendação" é mais usado por ser mais geral, não sendo realizada, necessariamente, nenhuma colaboração entre pessoas[6]. Já em 1996 o *Yahoo* utilizou sistemas de recomendação em uma de suas páginas, aplicando em larga escala[6], coisa que hoje é feita comumente por diversos sites e serviços.

É facilmente perceptível no cotidiano o uso de sistemas de recomendação em ambientes on-line. Ao usar a *Netflix*, sugestões para o usuário são oferecidas, baseadas nas atrações já assistidas e/ou avaliadas. Sites de compras como a *Amazon* também oferecem sugestões de produtos ao usuário baseado em visitas à página dos produtos ou no comportamento de outros usuários que compraram um mesmo produto. Também em redes sociais, como no

YouTube, são sugeridos vídeos baseados no histórico do internauta e nas suas avaliações, ou então no *Facebook*, que recomenda lista de pessoas que o usuário pode conhecer[7].

Em geral, sistemas de recomendação utilizam como informação a avaliação (*rating*) dada pelos usuários aos itens, podendo a avaliação estar expressa de diferentes maneiras[4]:

- Avaliações numéricas: O usuário avalia um item numa escala numérica, como no site da *Amazon*, onde o usuário dá uma nota de até 5 estrelas.
- Avaliações qualitativas: A avaliação é dada por frases definidas, como: "Concordo totalmente", "Concordo parcialmente", ...
- Avaliações binárias: O usuário seleciona se gostou ou não gostou do item, como a *Netflix*, atualmente, recebe as avaliações.
- Avaliação unária: A indicação se refere a se o usuário visualizou, comprou ou então avaliou o item positivamente.

Os algoritmos de recomendação utilizam uma matriz, chamada de matriz de avaliações (*ratings matrix*), usualmente representada desta forma:

Tabela 1: Típica matriz **R** de avaliações

	Item 1	Item 2	$\cdots$	Item m
Usuário 1	$r_{(1,1)}$		$\cdots$	
Usuário 2		$r_{(2,2)}$	$\cdots$	$r_{(2,m)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Usuário n			$\cdots$	$r_{(n,m)}$

Onde $r_{(i,j)}$ é a avaliação (*rating*) do usuário i dado ao item j. Em geral, os usuários não tiveram contato com todos os itens, então os itens não recebem avaliações de todos os usuários, produzindo então uma matriz esparsa (com grande quantidade de valores faltantes). Os algoritmos buscam, então, preencher a matriz de avaliações com previsões para os valores faltantes.

1.1 Técnicas de recomendação

Existem diferentes categorias de sistemas de recomendação, que podem ser classificados em: Filtragem baseada em conteúdo (*Content-based filtering*), filtragem colaborativa

(*Collaborative filtering*) e sistemas de recomendação híbridos (*Hybrid Recommender Systems*)[5].

1.1.1 Filtragem baseada em conteúdo

Os sistemas nesta categoria recomendam itens similares aos que o usuário gostou no passado[7]. Para isto é necessário utilizar informações das características de um produto[4] e comparar com o perfil do usuário, de acordo com itens já conhecidos pelo usuário. Considerando filmes como itens, se um usuário avaliou positivamente filmes do gênero de ação, então o sistema recomendará a este usuário filmes de ação. Por outro lado, a filtragem baseada em conteúdo não leva em conta a similaridade de preferência entre os usuários, mas apenas o histórico do usuário e as características dos itens[7].

Algumas das técnicas utilizadas neste tipo de filtragem são: TF/IDF (*Term Frequency Inverse Document*), *naive Bayes Classifier*, árvores de decisão ou redes neurais[1].

1.1.2 Filtragem colaborativa

Na filtragem colaborativa são recomendados itens de acordo com as avaliações de todos os usuários[5]. Para isto verifica-se a similaridade entre usuários (vizinhança), de acordo com suas avaliações passadas. Essa técnica é a mais utilizada para realizar recomendações[4]. Um exemplo simples seria: Se o usuário 1 comprou o item A, B e C, e o usuário 2 comprou os itens A e C, então recomenda-se o item B para o usuário 2.

1.1.3 Sistemas de recomendação híbridos

Os sistemas híbridos são uma combinação da filtragem baseada em conteúdo e filtragem colaborativa, buscando aproveitar as vantagens e eliminar as desvantagens das técnicas[4]. Cada uma das técnicas podem ser aplicadas de maneira separada, combinando os resultados, mas também pode ser construído um modelo com as duas abordagens unificadas[3].

2 Objetivos

Este trabalho tem os seguintes objetivos:

2.1 Objetivo geral

- Apresentar técnicas de sistemas de recomendação, executar algumas destas técnicas de filtragem colaborativa e compará-las.

2.2 Objetivos específicos

3 Materiais e Métodos

3.1 Conjunto de dados

Será utilizado um *dataset* disponível no site *grouplens*, disponível em `https://grouplens.org/datasets/movielens/1m/`. O conjunto de dados possui 1 000 209 avaliações de 3900 filmes dados por 6040 usuários[8], que se cadastraram no site *MovieLens* no ano de 2000.

Os usuários são representados pelo seu ID, que varia entre 1 e 6040 e os filmes possuem ID entre 1 e 3952. As avaliações têm formato numérico, de até 5 estrelas, com estrelas completas, tendo cada usuário avaliado ao menos 20 filmes.

3.2 Metodologia

Haverão um conjunto de usuários $U = \{u_1, u_2, \ldots, u_n\}$ e um conjunto de itens $I = \{i_1, i_2, \ldots, i_m\}$, assim como as notas dos usuários aos itens, que serão armazenadas na matrix $\mathbf{R}_{n \times m}$ de avaliações[9]. Logo, cada linha da matriz $\mathbf{R}$ representa um usuário e cada coluna, um item. Os algoritmos buscarão preencher os valores faltantes desta matriz, com valores na mesma escala das avaliações presentes na matriz[3].

3.2.1 Filtragem colaborativa baseada no item (item-based)

Este algoritmo busca recomendar itens similares aos bem avaliados pelo usuário. Desta forma será verificado, para cada par de itens, a sua similaridade, e a partir desta medida é prevista a avaliação do usuário para o item. A similaridade entre dois itens i e j pode ser medida pelo coeficiente de correlação de Pearson, definido da seguinte maneira[5]:

$$w_{i,j} = \frac{\sum_{u \in U} (r_{u,i} - \overline{r}_i)(r_{u,j} - \overline{r}_j)}{\sqrt{\sum_{u \in U} (r_{u,i} - \overline{r}_i)^2 \sum_{u \in U} (r_{u,j} - \overline{r}_j)^2}} \tag{3.1}$$

Sendo U o conjunto de usuários que avaliaram os dois itens, i e j,$r_{u,i}$ o rating dado

pelo usuário u ao item i e $\overline{r}_i$ o rating médio recebido pelo item i dado por todos os usuários que o avaliaram.

Alternativamente, a similaridade entre os itens i e j pode ser medida considerando os ratings recebidos pelos dois itens como vetores e calcular o cosseno entre estes vetores[10]:

$$w_{i,j} = \cos(\vec{r}_i, \vec{r}_j) = \frac{\vec{r}_i \cdot \vec{r}_j}{\|\vec{r}_i\| \times \|\vec{r}_i\|} = \frac{\sum_{u=1}^{n} r_{u,i} r_{u,j}}{\sqrt{\sum_{u=1}^{n} r_{u,i}^2 \sum_{u=1}^{n} r_{u,j}^2}} \tag{3.2}$$

A seguir, o *rating* do item i pelo usuário a pode ser previsto da seguinte forma[5]:

$$p_{a,i} = \frac{\sum_{j \in k} r_{a,i} - w_{i,j}}{\sum_{j \in k} |w_{i,j}|} \tag{3.3}$$

Sendo k o conjunto de itens avaliados pelo usuário a que são mais similares ao item i.

3.2.2 Filtragem colaborativa baseada no usuário (user-based)

Este algoritmo assume que usuários com preferência similar no passado terão preferências similares no futuro. Então os *ratings* não observados serão previstos a partir das avaliações de uma vizinhança e usuários com gostos similares[9]. São então encontrados os k vizinhos mais próximos de um usuário ou então todos os usuários que tenham pelo menos uma dada similaridade. O coeficiente de correlação de Pearson pode ser utilizado como medida de similaridade entre dois usuários a e u, definida da seguinte maneira[5]:

$$w_{a,u} = \frac{\sum_{i \in I} (r_{a,i} - \overline{r}_a)(r_{u,i} - \overline{r}_u)}{\sqrt{\sum_{i \in I} (r_{a,i} - \overline{r}_a)^2 \sum_{i \in I} (r_{u,i} - \overline{r}_u)^2}} \tag{3.4}$$

Sendo I o conjunto de itens avaliados pelos dois usuários, $r_{u,i}$ é o rating dado pelo usuário u ao item i e $\overline{r}_u$ é o rating médio do usuário u a todos os itens por ele avaliados.

Uma outra maneira de calcular a similaridade entre dois usuários é considerar os ratings de dois usuários como vetores num espaço m-dimensional, para, assim, encontrar o cosseno do ângulo entre estes vetores[5]:

$$w_{a,u} = \cos(\vec{r}_a, \vec{r}_u) = \frac{\vec{r}_a \cdot \vec{r}_u}{\|\vec{r}_a\| \times \|\vec{r}_u\|} = \frac{\sum_{i=1}^{m} r_{a,i} r_{u,i}}{\sqrt{\sum_{i=1}^{m} r_{a,i}^2 \sum_{i=1}^{m} r_{u,i}^2}} \tag{3.5}$$

Por fim, a predição da nota dada ao item i pelo usuário a é dada por:

$$p_{a,i} = \overline{r}_a + \frac{\sum_{u \in k} (r_{u,i} - \overline{r}_u) w_{a,u}}{\sum_{u \in k} |w_{a,u}|} \tag{3.6}$$

Sendo k a vizinhança do usuário a.

4 Análise dos Resultados

5 Conclusão

Referências

[1] ISINKAYE, F.; FOLAJIMI, Y.; OJOKOH, B. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, Elsevier, v. 16, n. 3, p. 261–273, 2015.

[2] MILD, A.; NATTER, M. Collaborative filtering or regression models for internet recommendation systems? *Journal of Targeting, Measurement and Analysis for marketing*, Springer, v. 10, n. 4, p. 304–313, 2002.

[3] TAKAHASHI, M. M.; JR, R. H. Estudo comparativo de algoritmos de recomendação. *USP. São Paulo*, 2015.

[4] SHAPIRA, B. et al. *Recommender systems handbook*. [S.l.]: Springer New York, 2011.

[5] MELVILLE, P.; SINDHWANI, V. Recommender systems. In: *Encyclopedia of machine learning*. [S.l.]: Springer, 2011. p. 829–838.

[6] REATEGUI, E. B.; CAZELLA, S. C. Sistemas de recomendaçao. In: *XXV Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2005. p. 306–348.

[7] GORAKALA, S. K.; USUELLI, M. *Building a recommendation system with R*. [S.l.]: Packt Publishing Ltd, 2015.

[8] HARPER, F. M.; KONSTAN, J. A. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, ACM, v. 5, n. 4, p. 19, 2016.

[9] HAHSLER, M. *recommenderlab: A framework for developing and testing recommendation algorithms*. [S.l.], 2015.

[10] SARWAR, B. et al. Item-based collaborative filtering recommendation algorithms. In: ACM. *Proceedings of the 10th international conference on World Wide Web*. [S.l.], 2001. p. 285–295.

ANEXO A – Título do primeiro anexo

ANEXO B – Título do segundo anexo