Avaliação de parâmetros para o algoritmo SVD regularizado

Cassiano H. da Silva[1], **Geovani S. Celebrim**[1]

[1]Departamento de Ciência da Computação
Universidade Federal Rural do Rio de Janeiro (UFRRJ)
26.020-740 – Nova Iguaçu – RJ – Brasil

{honoriocassiano, geovanicelebrim}@gmail.com

Resumo. *Motivados a tornar a experiência do usuário mais rica e interessante, os Sistemas de Recomendação tem como principal objetivo oferecer ao usuário um ambiente personalizado. Buscando obter técnicas mais robustas para essa personalização, os Sistemas de Recomendação baseados em modelo procuram otimizar o cálculo e armazenamento que é feito nos sistemas baseados em memória. Isso é feito criando-se um modelo onde o cálculo da recomendação seja mais rápido e escalável. Este trabalho propõe uma análise do algoritmo SVD regularizado, que busca fatores latentes que reúnem características de usuários e itens para realizar as predições. Ao final, é apresentado quais parâmetros melhor se adequaram para o cenário onde foram realizados os testes.*

1. Definição

Os Sistemas de Recomendação buscam, através de técnicas distintas, prover ao usuário uma experiência personalizada. Para isso, é realizado um estudo das características e relações entre itens e usuários para que um novo item possa ser recomendado a um usuário [Tail 2006]. Tais características nem sempre são explícitas – apesar de sabermos de sua existência – e não podem ser diretamente extraídas. Essas características são chamadas de variáveis latentes [Borsboom et al. 2003].

As técnicas de recomendação por filtragem colaborativa baseadas em modelo ganharam grande apelo especialmente após o concurso da Netflix Prize [Bennett and Lanning 2007], onde o algoritmo que fazia uso de fatores latentes [Koren et al. 2009] foi o que apresentou os melhores resultados. Nestes algoritmos, um modelo de predição de recomendação é proposto com treinamento utilizando um conjunto de dados previamente separado. Existem, na literatura, diversos tipos de modelos que apresentam bons resultados como *Support Vector Machines* (SVM) [Grčar et al. 2006] e *Singular Value Decomposition* (SVD) [Takács et al. 2009].

Este trabalho tem como principal objetivo explorar o modelo SVD, que é uma poderosa técnica de fatoração que busca encontrar um espaço de recursos de menor dimensão onde as novas características representam "conceitos" e o peso de cada conceito no contexto da coleção é computável [Shapira et al. 2011]. Com isso, o SVD permite obter automaticamente esses "conceitos" e utilizá-los como base para uma análise latente-semântica [Shapira et al. 2011], uma técnica popular de classificação de texto em Recuperação de Informação.

As predições através desse modelo podem ser obtidas por $\bar{R}_{ij} = P^T Q$. Onde as matrizes P e Q representa as características latentes dos usuários e itens, respectivamente. Durante o procedimento de treino, que almeja diminuir o erro e, assim, melhorar

as predições, deve-se realizar a minimização do erro quadrático apresentado na equação 1, que é equivalente ao problema dos mínimos quadrados [Helene 2006]. Nesta equação é calculado a diferença entre a nota real R e a nota predita $\bar{R}$ do usuário u para o item i. Fazendo o uso dos método do gradiente descendente, os valores para atualização de P e Q são mostrados nas equações 2 e 3, respectivamente, onde $lrate$ é a taxa de aprendizagem, e é o erro da predição e λ é a taxa de regularização.

$$\min \sum_{\substack{\forall u \in U \\ \forall i \in I}} (R_{ui} - \bar{R}_{ui})^2 \tag{1}$$

$$P_{iu} = P_{iu} + lrate * (e * Q_{ui} - \lambda * P_{iu}) \tag{2}$$

$$Q_{ui} = Q_{ui} + lrate * (e * P_{iu} - \lambda * Q_{ui}) \tag{3}$$

2. Metodologia

O conjunto de dados que foram utilizados trata-se de um *dataset* do *MovieLens*, disponível em `http://grouplens.org/datasets/movielens`. Inicialmente, os dados foram embaralhados e divididos em cinco conjuntos, cada um com 20% dos dados. Para cada variação do parâmetro, cinco experimentos foram realizados, onde em cada experimento, uma das cinco partes foi separada para teste enquanto as demais foram utilizadas para treino. O resultado final da avaliação do parâmetro é dado pela média aritmética dos resultados obtidos com os cinco experimentos.

O SVD regularizado (RSVD) [Funk 2006], que foi a técnica implementada neste trabalho, possui diversos parâmetros. São eles: λ, a taxa de regularização, k a dimensão de características a serem utilizadas e, por fim, $lrate$ que é a taxa de aprendizagem. Para que fosse encontrado o conjunto de parâmetros que produzia o melhor resultado para este conjunto de dados, o modelo foi submetido a variações destes parâmetros. Para cada parâmetro p, fixou-se os demais nos melhores valores até então encontrados e variou-se p para encontrar o valor que produzia o melhor resultado.

A seção seguinte apresenta os resultados dos experimentos realizados para cada variação dos parâmetros, permitindo assim, observar quais parâmetros se adequam melhor a esse cenário. Cabe ressaltar que encontrar o melhor conjunto de parâmetros é uma tarefa custosa, pois é realizada através de testes empíricos e esses testes podem ser demorados, tornando esse processo muito custoso e, dependendo do caso até inviável.

3. Experimentos e Resultados

Esta seção mostra os resultados obtidos com o RSVD, submetido às diversas variações de seus parâmetros. Para analisar a acurácia das técnicas de similaridade, foram utilizadas duas medidas comuns para avaliar o desempenho das recomendações. São elas: *Mean Absolute Error* (MAE), dada pela equação 4 e *Root Mean Squared Error* (RMSE), dada pela equação 5. O critério de parada utilizado em todos os testes, foi dado por uma taxa de melhora inferior a 0.0001 no RMSE ou por um número máximo de iterações $i = 250$.

$$MAE(f) = \frac{\sum_{r_{ui} \in R_{test}} |f(u,i) - r_{ui}|}{|R_{test}|} \tag{4}$$

$$RMSE(f) = \sqrt{\frac{\sum_{r_{ui} \in R_{test}} (|f(u,i) - r_{ui}|)^2}{|R_{test}|}} \tag{5}$$

A figura 1 apresenta o processo de melhora das predições do RSVD. A melhora pode ser observada com a minimização do MAE, no gráfico à esquerda ou do RMSE, no gráfico à direita. Através dos gráficos, é possível analisar que o erro, tanto do treino quanto do teste decresce rapidamente até que começa a se estabilizar. A melhora se dá através dos reajustes das matrizes de características latentes P e Q. Já a estabilização se dá quando o método não consegue mais reajustar essa matriz, de modo que melhore suas predições. É importante verificar que o erro do conjunto de teste se estabiliza de forma mais rápida enquanto o erro do conjunto de treino continua diminuindo. Isso mostra que de fato o modelo está buscando se adequar ao conjunto de dados utilizados para treinamento, mesmo que isso já não interfira tanto nos resultados das predições do conjunto de teste.

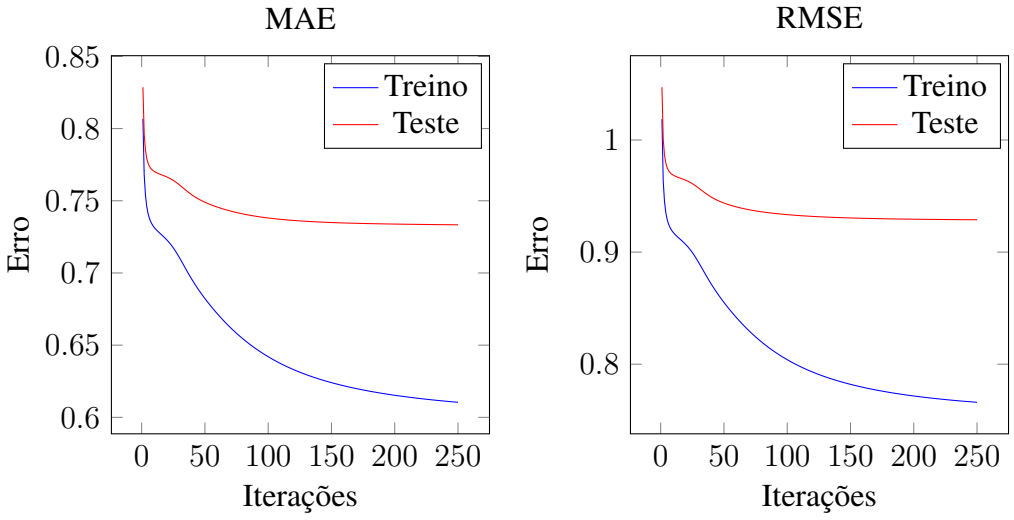

Figura 1. Melhoria da predição.

A figura 2 mostra o gráfico que apresenta os erros MAE e RMSE do RSVD, variando-se o λ, que é a taxa de regularização responsável por "controlar"as atualizações das matrizes de variáveis latentes. Para este teste, o parâmetro analisado iniciou-se com o valor 0.05 e foi incrementado em 0.01, para cada caso de teste, até atingir o valor máximo, que foi definido como sendo 0.15. Através da análise do gráfico é possível inferir que a melhor configuração para o λ, fixando os demais parâmetros é $\lambda = 0.11$, chegando a apresentar um MAE = 0.727 e RMSE = 0.915.

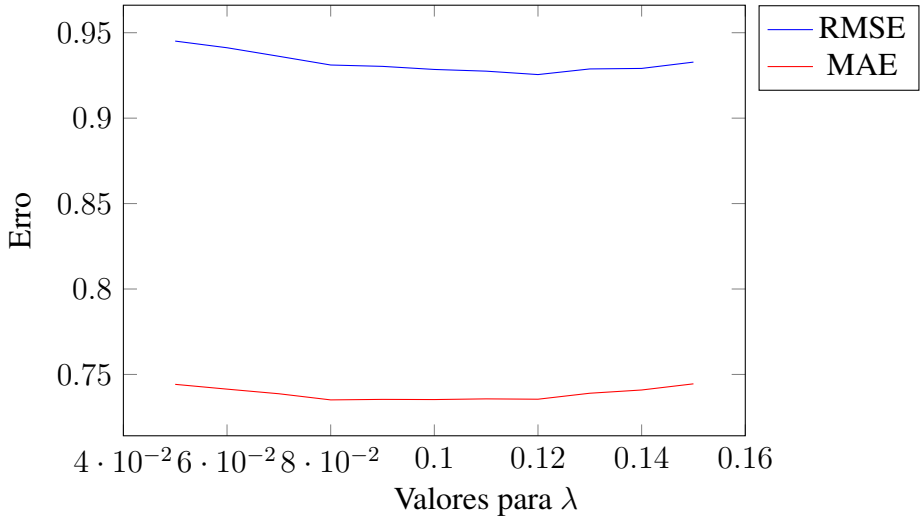

Figura 2. Relação entre λ e o erro da predição.

A figura 3 mostra também o gráfico do MAE e do RMSE, porém variando-se o $lrate$. Esse parâmetro controla a taxa de "aprendizagem"do RSVD. Pôde-se observar que esse é um parâmetro essencial para o bom funcionamento do algoritmo, afinal, o gráfico em momento algum ficou estável. É possível observar que ao iniciar os testes o modelo apresentou o melhor resultado, com MAE = 0.728 e RMSE = 0.922. A partir deste valor, o resultado piora quase que proporcionalmente ao incremento de $lrate$ e pode resultar em *overflow* dos valores das matrizes. O melhor valor do parâmetro para estes testes é 0.0066.

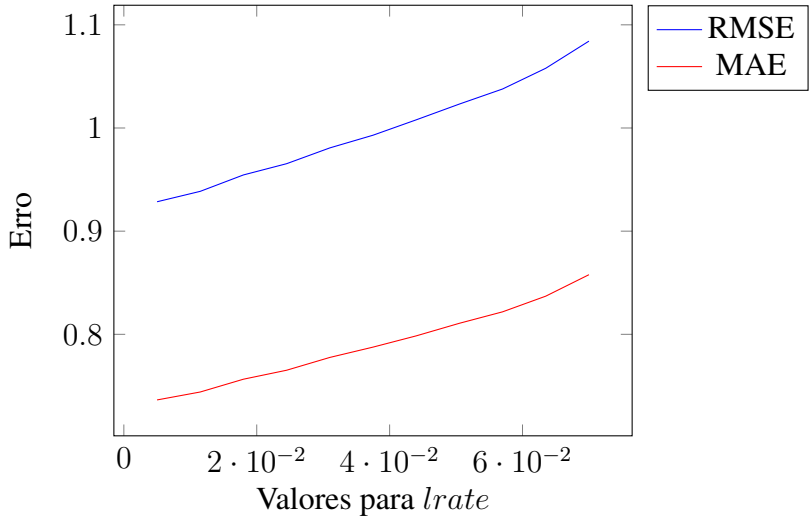

Figura 3. Relação entre $lrate$ e o erro da predição.

Por fim, a figura 4 apresenta o gráfico do MAE e RMSE, variando-se o k. Esse parâmetro representa a dimensão de características ou variáveis latentes que serão consideradas na predição. Os resultados mostram que a partir de um determinado valor, a variação de k não apresenta grandes variações no resultado. Pode-se observar que aumentar k indefinidamente, apesar de considerar mais características para a predição, não

necessariamente irá melhorar os resultados, inclusive essa alteração implicou em uma piora para esse cenário. Neste caso de teste, o parâmetro k iniciou-se com o valor 1 e foi até 100, em intervalos unitários. Os melhores resultados obtidos foram com $k = 30$, alcançando MAE = 0.725 e RMSE = 0.912.

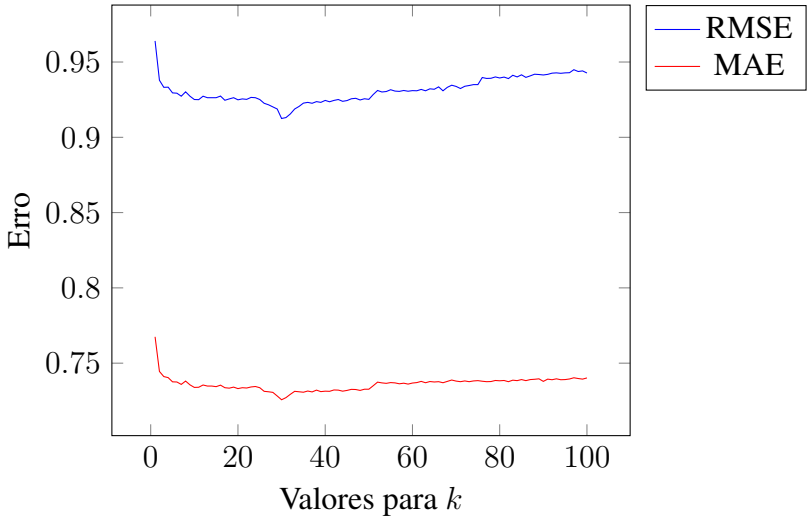

Figura 4. Relação entre k e o erro da predição.

4. Conclusão

Este trabalho teve como principal finalidade realizar uma análise do impacto dos parâmetros sobre as predições do algoritmo RSVD, método utilizado em sistemas de recomendação por filtragem colaborativa baseado em modelo. Utilizar o método baseado em modelo apresenta algumas vantagens se comparado aos métodos baseados em memória, uma delas é a simplicidade dos algoritmos baseados em modelo e seu melhor desempenho, na média. Outra vantagem é a escalabilidade fornecida por essa técnica, resolvendo de forma mais adequada o problema da inserção de um novo usuário ou item.

Analisando os resultados obtidos com a variação dos parâmetros do RSVD, conclui-se que um bom ajuste para os parâmetros para este cenário é $\lambda = 0.11$, $k = 30$ e $lrate = 0.0066$, chegando a obter RMSE = 0.912 e MAE = 0.725. Durante os testes, foi percebido também a sensibilidade que o RSVD possui. Pequenas alterações na configuração dos seus parâmetros podem acarretar em mudanças consideráveis no resultado. Outro fator observado é que muitas vezes o resultado final pode ser parecido, no entanto, a velocidade de convergência é diferente. Isso pode ser relevante quando o tempo é um fator limitante, se fazendo necessário estudar até que ponto vale a pena esperar várias iterações por uma melhora baixa.

Por fim, conclui-se que identificar os melhores parâmetros para o cenário do problema é fundamental, afinal, esses parâmetros mudam de acordo com as características do problema. Esse passo é essencial para que se obtenha resultados satisfatórios, pois como foi mostrado, os resultados das predições estão intimamente relacionados aos parâmetros do RSVD.

Referências

Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35.

Borsboom, D., Mellenbergh, G. J., and Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review*, 110(2):203.

Funk, S. (2006). Netflix update: Try this at home (december 2006).

Grčar, M., Fortuna, B., Mladenič, D., and Grobelnik, M. (2006). knn versus svm in the collaborative filtering framework. In *Data Science and Classification*, pages 251–260. Springer.

Helene, O. (2006). *Metodos dos Minimos Quadrados*. Editora Livraria da Física.

Koren, Y., Bell, R., Volinsky, C., et al. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Shapira, B., Ricci, F., Kantor, P. B., and Rokach, L. (2011). Recommender systems handbook.

Tail, L. (2006). Why the future of business is selling less of more.

Takács, G., Pilászy, I., Németh, B., and Tikk, D. (2009). Scalable collaborative filtering approaches for large recommender systems. *Journal of machine learning research*, 10(Mar):623–656.