Abordagem comparativa de técnicas de mineração de dados em pré-processamento de sistemas de recomendação.

Santos, L.*, Ariza N, F.*, Loiola, M.B.*

*Universidade Federal do ABC, Santo André, Brasil

e-mail: laurindo.santos@ufabc.edu.br, francisco.ariza@ufabc.edu.br

Abstract – *This paper aims to present preliminary results of some popular data mining techniques applied to recommender systems. By using a small sample of a films' database publicly available on the Internet, a comparative study among some popular data mining techniques is presented, as well as which of them can offer a clear understanding of data sets with similar characteristics.*

Palavras-chave: *Recommender Systems, Data Mining, Machine Learning, Bayesian Regression, KNN.*

Introdução

O intenso crescimento da quantidade de dados digitais gerados em meios como internet, games, *streaming* e consumo massivo de serviços on-line disponíveis geraram um vasto ambiente para exploração analítica, cujo acesso, processamento e transformação demandam alto poder computacional para que se obtenham resultados de pesquisa favoráveis; não raro, não oferecem aderência adequada aos argumentos de busca dos consulentes.

Alguns dos mais importantes buscadores de informação, como Google, Yahoo, Bing e outros, criaram seus próprios modelos de armazenamento e referência, mas a priorização e personalização de informações, quando um sistema mapeia o conteúdo disponível face aos interesses e preferências de quem o consulta, nem sempre se fazem presentes. Para empresas que oferecem serviço massivo de *streaming* de vídeo e sites de *e-commerce*, por exemplo, tão importante quanto os resultados produzidos através de uma busca semântica das palavras digitadas, são as recomendações de itens similares ao que está sendo buscado. Para isso, utilizam seus sistemas de recomendação.

Sistemas de recomendação são, sistemas de filtragem de dados que lidam com o problema de sobrecarga de informação [1] pela extração de fragmentos vitais de informação a partir de grande quantidade de dados gerados dinamicamente em função das preferências ou interesses dos consulentes ou, ainda, de seu comportamento observado em relação aos itens buscados [2]. Os sistemas de recomendação beneficiam tanto os provedores de serviços quanto os seus utilizadores [3], ao reduzirem os custos das transações de busca, seleção e oferta online de itens de resposta [4].

O uso de técnicas que produzam recomendações relevantes, com acurácia crescente, impõe-se e se justifica. Modelos de aprendizado de máquina e algoritmos de mineração de dados são largamente empregados nos processos de análise de requisições e oferta de resultados, considerando modelos de filtragem, classificação, agrupamento e associação. KNN (K-Nearest Neighbors) [5] é uma técnica de mineração de dados aplicável com esse propósito e que opera buscando semelhanças entre itens (os k vizinhos mais próximos). Filtragem colaborativa é técnica de predição aplicável a conteúdos que não possam ser adequadamente definidos por metadados, tais como música ou filmes [6]. Conjugadas, realizam recomendações associando usuários com interesses e preferências comuns ou aproximadas, a vizinhança, contidos em uma base de dados (matriz usuário-item) de itens preferidos pelos usuários.

O objetivo deste trabalho é, portanto, realizar uma experimentação da técnica KNN e um estudo comparativo entre essa técnica e outros algoritmos de regressão ou classificação, realizando uma abordagem comparativa que demonstre a efetividade e acurácia de cada uma das técnicas para contextos similares ao utilizado neste estudo, que serve como referência para a aplicação dessas técnicas como parte integrante dos sistemas de recomendação.

Metodologia

Neste trabalho foram comparadas as técnicas KNN, SVM (Support Vector Machine) [7], Regressão Bayesiana [8] e Árvore de Decisão [9].

Por se tratarem, basicamente, de modelos de análises exploratórias, estatísticas e matemáticas, os softwares de apoio utilizados foram MS Excel, para manipulação de dados no formato CSV, e Jupyter Notebook, para execução dos modelos utilizando linguagem Python, por possuir bibliotecas como *sklearn*, para trabalhar as técnicas estatísticas, e *numpy*, para representações matemáticas.

Os dados para estudo foram obtidos no Internet Movies Data Base (IMDB), em seu site da Web (http://www.imdb.com/interfaces); trata-se de base de dados de filmes, composta por atributos que caracterizam gênero, principais envolvidos e outras, enriquecidos com dados de redes sociais do Facebook, com opiniões emitidas por usuários sobre cada um dos filmes. Utilizou-se amostra de aproximadamente 5000 registros com 28 colunas, das quais apenas 15 variáveis numéricas e contínuas foram utilizadas na exploração das técnicas. A escolha da quantidade de registros deu-se pela eventual necessidade de corroborar manualmente conclusões obtidas através dos modelos.

O primeiro passo da investigação consistiu na análise dos dados, visando identificar parâmetros de entrada relevantes para a execução das técnicas para os casos estudados; campos em branco foram substituídos por 0, por se tratarem de valores nulos não preenchidos no CSV. Seguiu-se, então, a seleção dos campos numéricos contínuos, num total de 15.

Um passo significativo para entendimento dos dados foi realizar uma validação cruzada entre eles, buscando identificar a influência cruzada entre atributos. É importante avaliar se campos altamente correlacionados podem apresentar a mesma informação: dois ou mais campos que o façam podem acrescentar peso indevido a determinado atributo, distorcendo os resultados. A validação cruzada busca mostrar o quanto os dados opiniões, identificados no parágrafo anterior, estão correlacionados entre si e aos dados filmográficos. Exemplarmente, percebe-se o quanto o campo *orçamento* influencia o campo *quantidade de "likes"* do filme. A Figura 1 ajuda a entender como é possível identificar essa influência.

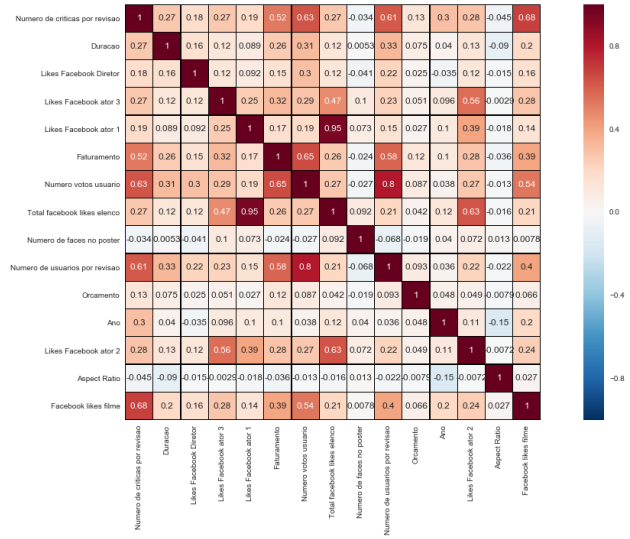

Figura 1. Validação cruzada de atributos

Buscou-se entender como estava representada a variância da população de atributos e constatou-se, através da representação gráfica, que apenas 1/3 dela representa aproximadamente 70% da variância total encontrada na amostra. Com efeito, a Figura 2 mostra que apenas 5 dos atributos representam 70% da variância, que significa que a média da população está bem representada nesses componentes. O estudo comparativo das técnicas tomou esse número como parâmetro sempre que necessário.

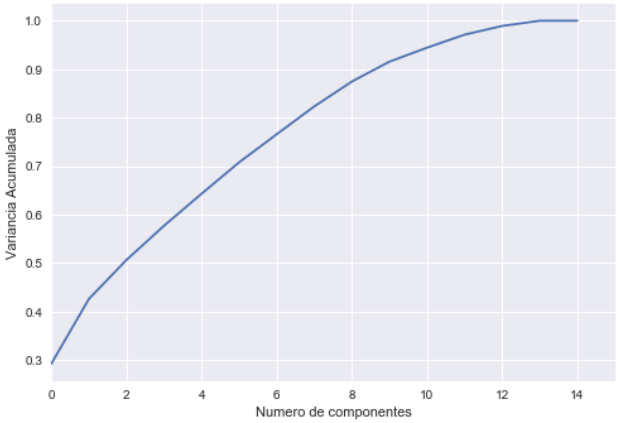

Figura 2. Gráfico de variância

A preparação de dados de testes e de validação tomou 25% deles para testes e os 75% restantes para validação dos modelos. Os dados apresentados na Tabela 1 apresentam os resultados obtidos pelas técnicas considerando esses percentuais.

Para cada uma das técnicas estudadas, os resultados obtidos, a distribuição e os resíduos identificados são apresentados em gráficos de dispersão. Resíduos caracterizam-se pelos pontos mais esparsos, indicando opiniões ("*ratings*") pouco aderentes às demais. As regiões mais condensadas indicam avaliações mais coesas, mais aderentes e, portanto, passíveis de utilização em recomendações mais acuradas. As taxas de acerto para dados de treinamento e dados de validação são apresentados, de forma unificada e para todas as técnicas, na sessão Discussão.

Técnicas e Resultados

A primeira técnica explorada foi o SVM, técnica que vem recebendo crescente atenção da comunidade de aprendizado de máquina e está embasada pela teoria de aprendizado estatístico, desenvolvida por Vapnik [9]. SVM é um conceito na ciência da computação para um conjunto de métodos do aprendizado supervisionado que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão. Os resultados

obtidos com a utilização de SVM estão apresentados na Figura 3.

Figura 3. Gráfico de dispersão SVM

O modelo KNN [5] foi proposto por Fukunaga e Narendra em 1975 e trata-se de técnica não paramétrica usada para regressão e classificação, principal interesse neste estudo. Os resultados obtidos estão plotadas na Figura 4 e as taxas de acertos obtidas com dados de testes e validação estão apresentadas na Tabela 1.

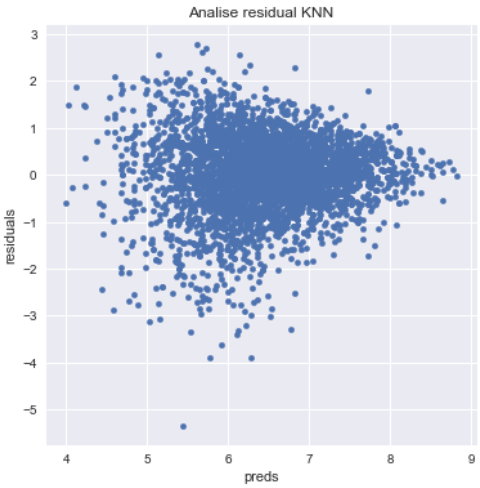

Figura 4. Gráfico de dispersão KNN

A Regressão Bayesiana [8] tem base em regressão linear, utilizando a inferência bayesiana para lidar com as incertezas que são modificadas periodicamente após observações de novos dados ou resultados. A inferência bayesiana é de natureza estatística e descreve as incertezas sobre quantidades invisíveis de forma probabilística. A regra de Bayes pode ser aplicada de forma iterativa: probabilidades obtidas após alguns eventos passam a ser consideradas probabilidades prévias para novos eventos observados. Os resultados obtidos no emprego desta técnica são vistos na Figura 5.

Figura 5. Gráfico de dispersão regressão bayesiana

Muito úteis em processos de mineração de dados de grandes bases de dados, Árvores de Decisão são abordagens simplificadas do conhecimento humano, largamente utilizadas na construção de algoritmos de classificação. Baseiam-se em estimativas e probabilidades associadas a eventos em que cada resultado é ponderado pela probabilidade associada a ele. O resultado é somado e o valor esperado de cada sequência de eventos é, assim, determinado.

Árvores de Decisão, para este caso, foram baseadas em um modelo linear, assim como as árvores de decisão simples fazem uma decisão ramificada baseada em uma série de valores dados como entrada [11]. Os resultados obtidos foram plotados no padrão abaixo e calculadas as taxas de acertos para dados de testes e validação:

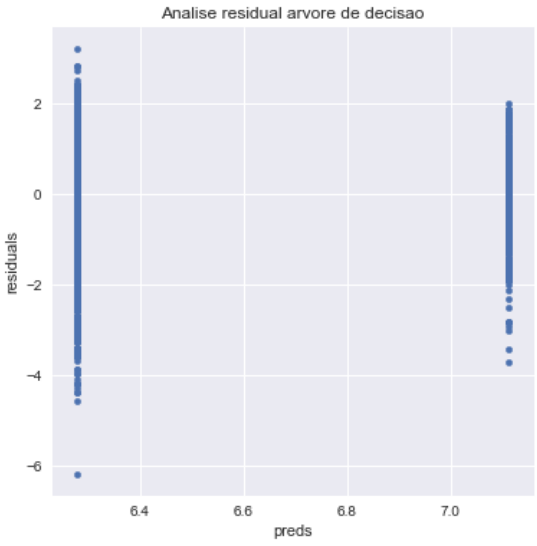

Figura 6. Gráfico de dispersão árvore de decisão

Discussão

Com os resultados obtidos, tendo em conta o contexto dos dados em que as técnicas foram aplicadas, é importante ressaltar que a mudança de qualquer dos atributos pode resultar em alterações naqueles resultados. Essa análise, porém, não fez parte do estudo, bem como não houve a pretensão de encontrar técnicas aplicáveis a qualquer contexto.

As técnicas estudadas e implementadas neste trabalho servirão de base para o pré-processamento de dados em estudos posteriores em sistemas de recomendação. Nesse aspecto, as técnicas utilizadas serão somente uma parte dos modelos a serem estudados e, assim sendo, pode ser necessário o estudo e uso de outras técnicas para que se obtenham melhores resultados. Também não significa que as melhores taxas de acerto obtidas aqui serão os algoritmos que terão maior assertividade após serem combinados com outras partes dos sistemas de recomendação.

Executadas todas as técnicas mencionadas, tendo como base o KNN, os resultados obtidos estão descritos na Tabela 1 abaixo:

Tabela 1.Taxas de acerto

Técnica	Taxa de Acerto Validação	Taxa de Acerto Treinamento
Knn	93,06%	87,67%
Regressão Bayesian	99,96%	84,54%
Árvore de Decisão	93,44%	83,75%
SVM	94,92%	86,08%

Muito importante considerar que a análise inicial dos dados, da mesma forma que a avaliação cruzada, que eliminou as influências entre atributos, permitiu a obtenção de uma base de dados para avaliação dotada de forte característica de homogeneidade.

Pode-se observar que os resultados são bastante similares em alguns casos, fruto dessa homogeneidade, destacando-se a Regressão Bayesiana quanto aos dados de validação. Esses resultados não indicam, entretanto que a Regressão Bayesiana deva ser a única técnica contemplada no estudo de sistemas de recomendação. Deve-se compreender que, devido à complexidade dos sistemas de recomendação, e considerando também importantes outros aspectos não considerados neste estudo, como, por exemplo, avaliação de desempenho *versus* volume de dados, é importante elencar mais de uma dessas técnicas no contexto de análises mais aprofundadas.

Conclusões

As comparações subsidiam considerar a convergência de obtenção de resultados pelas técnicas avaliadas em situações de bases de dados caracterizadas por atributos qualitativamente homogêneas. Surge, neste caso, a Regressão Bayesiana como a técnica de maior destaque, superando em mais de 5% a segunda colocada, servindo como respaldo para que se prossiga utilizando essa técnica para realização dos sistemas de recomendação e possivelmente considerar as demais na ordem de acurácia obtida em cada uma delas na fase de validação,

Referências

[1]Konstan JA, Riedl J. Recommender systems: from algorithms to user experience. User Model User-Adapt Interact 2012.

[2]Pan C, Li W. Research paper recommendation with topic analysis. In Computer Design and Applications IEEE 2010.

[3]Pu P, Chen L, Hu R. A user-centric evaluation framework for recommender systems. In: Proceedings of the fifth ACM conference on Recommender Systems (RecSys'11), ACM, New York, NY, USA; 2011.

[4]Hu R, Pu P. Potential acceptance issues of personality-ASED recommender systems. In: Proceedings of ACM conference on recommender systems (RecSys'09), New York City, NY, USA; October 2009.

[5]Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression".The American Statistician. 46 (3): 175–185.

[6]Isinkaye F.O., Folajimi Y.O., Ojokoh B.A. Recommendation Systems: Principles, methods and evaluation. Egyptian Informatics Journal 2015

[7]Celma, Ò. (2010), The Recommendation Problem, in Music Recommendation and Discovery, Berlin Heidelberg: Springer-Verlag, pp. 15-41.

[8]T. Mitchell. Machine Learning.McGraw Hill, 1997.

[9]V. N. Vapnik. The nature of Statistical learning theory. Springer-Verlag, New York, 1995.

[10]BOX, G. E.; TIAO, G. C. Bayesian inference in statistical analysis. New York: John Wiley & Sons, 1973. 360p.

[11]Alfred V. Aho, (1983). Data structures and algorithms, por John E. Hopcroft, Jeffrey D. Ullman