PROGRAMAÇÃO E BANCO DE DADOS

Priscila Gonçalves

Aplicar metodologias de *Data Mining* (mineração de dados)

Objetivos de aprendizagem

Ao final deste texto, você deve apresentar os seguintes aprendizados:

- Reconhecer o conceito de *Data Mining*.
- Identificar as principais técnicas de *Data Mining*.
- Utilizar as metodologias de *Data Mining*.

Introdução

O processo em que grandes quantidades de dados são explorados com o objetivo de identificar padrões, relacionamentos, conhecimentos é denominado *Data Mining* (em português, mineração de dados) e tem cada vez mais importância para o mercado, para os negócios e mesmo para pesquisas científicas, que têm interesse e necessidade de analisar e organizar a quantidade enorme de dados que produzem.

Por isso, neste capítulo, você aprenderá a reconhecer o conceito de *Data Mining*, verá como identificar as principais técnicas e utilizar as metodologias de *Data Mining*.

O que é *Data Mining*?

Com *Data Mining* (em português, mineração de dados), é possível descobrir informações de grande valor, principalmente para ajudar nas tomadas de decisões. A mineração de dados utiliza como base para seus trabalhos experimentos de áreas como estatística, inteligência artificial, máquina de estado e banco de dados para construir seu modelo.

A mineração de dados está relacionada, também, às áreas da inteligência artificial que são chamadas de descoberta de conhecimento e aprendizagem

de máquina. O termo "mineração de dados" está relacionado aos estágios de descoberta do processo de KDD (*Knowledge Discovery in Databases*), que "é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O termo "não trivial" diz respeito à complexidade existente na execução e manutenção dos processos de KDD; o termo "interativo" representa a relevância de ter um elemento que controle o processo; o termo "iterativo" indica a possibilidade de repetições em qualquer uma das etapas do processo; e o "conhecimento útil" é a há indicação de que o objetivo foi alcançado. A fase mais importante do processo de KDD é a mineração de dados aplicada, pois é nela que são utilizados algoritmos e determinada técnica que tem como objetivo elaborar um modelo para representar um conjunto de dados. Essa fase baseia-se em técnicas de estatística, inteligência artificial, computação paralela e máquina de estado, construindo um histórico de pesquisas relacionadas a essas áreas. Além disso, busca padrões, relacionamentos entre dados, anomalias e regras, tendo como objetivo encontrar informações ocultas que sejam relevantes para tomadas de decisões.

Dentre as características mais importantes da mineração de dados, está o grande volume de dados e a capacidade de mudança de escala com relação ao tamanho dos dados. Algoritmos têm a capacidade de mudança de escala, mas a mineração é muito mais do que aplicar algoritmos, pois, geralmente, os dados contém ruído ou estão incompletos, sendo provável que padrões sejam perdidos e a confiabilidade, baixa. Logo, o analista precisa tomar a decisão sobre quais tipos de algoritmos de mineração serão necessários, aplicando-os em um conjunto de amostra de dados específico, sintetizando os resultados, aplicando ferramentas de apoio à decisão e mineração, iterando o processo.

As principais técnicas de *Data Mining*

Dentre os tipos de dados que podem ser minerados, utilizam-se técnicas diferentes de mineração. Esse processo de definição e criação do modelo que será utilizado é a maior parte do processo, na qual deverão ser incluídas as perguntas sobre os dados e deverá constar um modelo de respostas para as perguntas feitas; a partir disso, será implantado o modelo propriamente dito.

Vários algoritmos e técnicas podem ser utilizados nesse processo. Podemos citar os seguintes algoritmos: associação, itens frequentes, *clustering*, árvores de decisão, classificação bayesiana, mineração por redes neurais.

Mineração por grupo de associação

A técnica de mineração por associação tem por objetivo identificar o relacionamento de itens que, em um específico conjunto de dados, sejam mais frequentes. Normalmente, o volume de dados que envolvem esse tipo de mineração é extenso e, diante dessa premissa, torna-se necessária a utilização de algoritmos que sejam mais rápidos e eficientes.

A seguir, veja um exemplo de mineração de dados por associação:

> Regra 1: SE idade > 25 AND graduação completa = sim ENTÃO fazer mestrado = sim
>
> Regra 2: SE idade <= 25 AND graduação completa = não ENTÃO fazer mestrado = não

Mineração de itens frequentes

Esta técnica, geralmente, é visualizada em duas etapas: na primeira delas, um conjunto de itens frequentes é desenvolvido e há um valor mínimo de frequência a ser respeitado. Após essa etapa, regras de associação devem ser geradas pela mineração desse conjunto de itens. A fim de que os resultados sejam válidos, para cada regra produzida, deverão ser utilizados conceitos de confiança e suporte. Os conceitos referentes a suporte são referentes ao percentual de registros que se enquadram na regra, e os conceitos de confiança medem o percentual de registros de uma forma específica para a regra.

O algoritmo mais utilizado para a estratégia da mineração de itens frequentes é o *Apriori*, no qual são envolvidas técnicas de *hash*, particionamento, redução de transações e segmentação.

Mineração por *clustering*

A técnica de *clustering* tem como objetivo identificar e aproximar dados semelhantes. Trata-se de uma coleção de registros semelhantes entre si, mas diferentes de registros em demais agrupamentos. Essa técnica não

pretende classificar, estimar ou predizer o valor de qualquer variável, apenas pretende identificar grupos de dados similares. Existem algumas tarefas para as quais essa técnica é bastante utilizada: pesquisa mercadológica, reconhecimento de padrões, processamento de imagens, análise de dados, taxonomia de plantas e também de animais, segmentação mercadológica, pesquisas geográficas, detecção de fraudes, classificar documentos presentes na *web*, etc.

A Figura 1, a seguir, representa a mineração de dados por *clustering*.

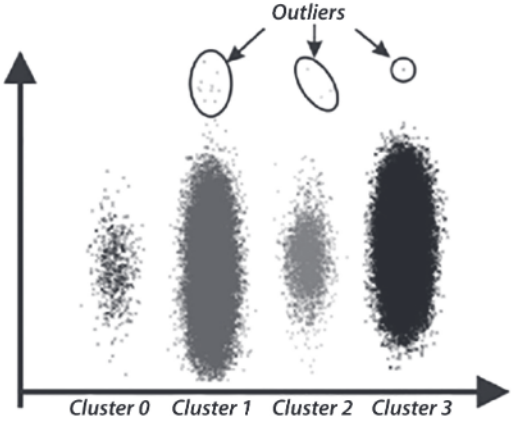

Figura 1. Mineração de dados por *clustering*.
Fonte: Salvador, Cunha e Corrêa (2009, p. 452).

Mineração por árvores de decisão

A técnica de mineração por árvores de decisão faz muito sucesso devido ao fato de não necessitar de parâmetros de configuração (o que a torna bastante simples) e por ter um alto grau de assertividade. Geralmente, é utilizada em categorizações ou previsões de dados. Árvores de decisão são formadas a partir de um conjunto de regras de classificação, em que cada caminho da raiz até uma folha representa uma dessas regras. Mesmo sendo uma técnica muito poderosa, faz-se necessário uma análise detalhada dos dados que deverão ser

utilizados, garantindo, assim, os melhores resultados. Geralmente, a árvore de decisão é definida a fim de que, para cada observação referente à base de dados, haja um e somente um caminho da raiz até a folha (Figura 2).

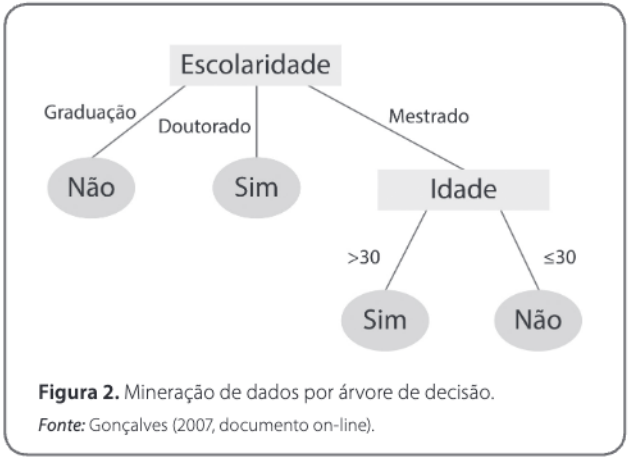

Figura 2. Mineração de dados por árvore de decisão.
Fonte: Gonçalves (2007, documento on-line).

Mineração por classificação bayesiana

A técnica de mineração por classificação bayesiana é tida como uma técnica estatística e baseia-se no teorema de Thomas Bayes, segundo o qual é possível encontrar a probabilidade de um determinado evento ocorrer diante da probabilidade de outro evento já ter ocorrido:

Probabilidade (Y dado X) = Probabilidade (X e Y) / Probabilidade (X).

Esse tipo de algoritmo, Naive Bayes, obtém resultados compatíveis com os resultados das árvores de decisões e, por ser simples e ter um alto poder de prever, é um dos tipos mais utilizados. Esse algoritmo parte do princípio de que não haja relação de dependência entre os atributos, mas nem sempre isso ocorre.

Mineração por redes neurais

A técnica de mineração por redes neurais tem sua origem na psicologia e na neurobiologia e consiste em simular o comportamento dos neurônios. Pode ser

vista como um conjunto de entradas e saídas (assim como ocorre nos neurônios) que são conectadas por camadas intermediárias e na qual cada ligação tem um valor associado. É uma técnica que precisa de um grande período de treinamento, ajustes de parâmetros. É difícil de interpretar e também não é possível identificar de forma clara e precisa a relação entre a entrada e a saída. Porém, essas redes neurais conseguem trabalhar de maneira que não tenham problemas de valores errados e podem identificar padrões para os quais nunca foram treinadas.

Na Figura 4, temos um exemplo do algoritmo *backpropagation*, que é um dos mais conhecidos nas redes neurais e aprende a partir da correção de erros.

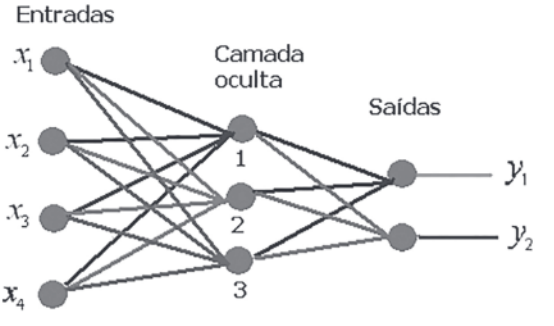

Figura 3. Exemplo de algoritmo *backpropagation*.
Fonte: Lanhellas (2013, documento on-line).

Fique atento

Uma árvore de decisão é uma representação de uma coleção de regras de classificação. Cada nó interno da árvore é rotulado com um atributo previsor, que frequentemente é chamado de atributo de divisão. É com base nas condições desse atributo que os dados são divididos (RAMAKRISHNAN; GEHRKE, 2013).

Utilização das metodologias de mineração

Todos os dias, empresas trabalham com uma enorme quantidade de dados, seja com informações cadastrais, preferências de consumidores, interações em redes sociais e transações com clientes. Quando esses dados são organizados e analisados por metodologias de *Data Mining*, podem garantir o sucesso das empresas, principalmente na tomada de decisões. Com a utilização das metodologias de mineração, é possível fazer correlações, desvendar tendências e verificar a existência de padrões; dessa forma, consegue-se abstrair o conhecimento necessário para alavancar os negócios e tomar as decisões corretas.

Dentre as metodologias, pode-se citar CRISP-DM (*Cross Industry Standard Process for Data Mining*), que se trata de uma metodologia elaborada especificamente para processos de mineração de dados. O método CRISP-DM é dividido em 6 partes:

- Entendimento do negócio: o profissional deve procurar compreender o problema a ser solucionado, buscando entender como o problema afeta a empresa e quais são os objetivos a serem alcançados.
- Compreensão de dados: tem por objetivo verificar, descrever e organizar os dados.
- Preparação dos dados: após sua definição, organização e verificação, os dados deverão ser conduzidos pelo profissional de forma técnica, definindo, inclusive, o formato necessário para analisá-los.
- Modelagem: as técnicas de mineração são selecionadas e aplicadas de acordo com os objetivos a serem alcançados.
- Avaliação: etapa na qual ocorre o acompanhamento dos resultados e a avaliação da aplicabilidade dos conhecimentos adquiridos.
- Desenvolvimento: todo o conhecimento obtido pela mineração será aplicado de uma forma mais prática, apresentando uma entrega aplicável ao cliente, em que o mesmo possa facilmente verificar os resultados concretos obtidos a partir da análise de dados.

Na Figura 4, são apresentadas as 6 etapas da metodologia CRISP-DM.

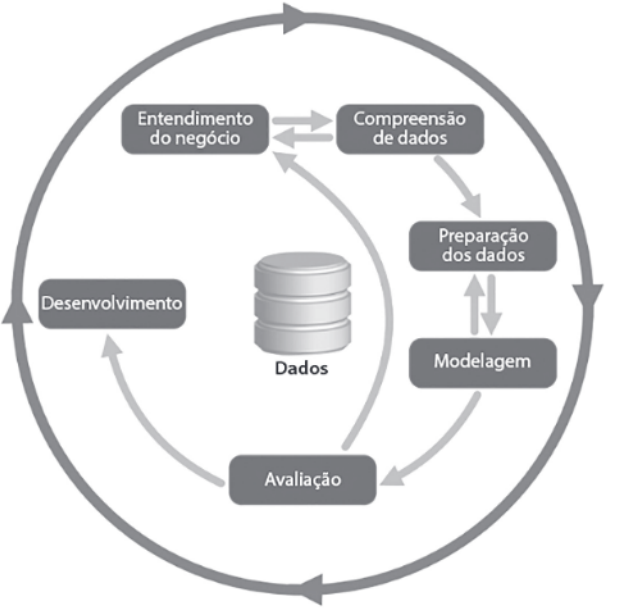

Figura 4. Etapas da metodologia CRISP-DM.
Fonte: Hekima (2016, documento on-line).

Existem, também, outras aplicações de mineração de dados. Pode-se citar:

- *Basket analysis*: é realizada uma análise de afinidade e pode ser aplicada a vários objetos, identificando combinações de itens com foco no padrão de compras de consumidores. Metodologia muito aplicada em *e-commerce*.
- Análises preditivas: a metodologia ajuda a prever quando os clientes irão realizar novas compras, para que as empresas possam realizar campanhas de marketing, organizar estoques e traçar cenários.
- Monitoramento de redes sociais: são verificados dados por meio da interação com conteúdos que instiguem o consumidor a "dizer", por meio de curtidas, comentários e compartilhamentos, o que

pensam sobre determinados produtos, marcas e até mesmo tipo de atendimentos.

- Mineração de dados e OLAP (*On-Line Analytical Processing*): conceito que engloba análises rápidas de dados multidimensionais compartilhados, complementando a mineração de dados. Os sistemas de OLAP são ideais para alocação de custos, análises de séries temporais, indexação de dados e análises "*what-if*". Trata-se de uma forma de analisar negócios e empresas naturalmente.

Saiba mais

Acesse o site a seguir e saiba mais a respeito das consultas das aplicações de *Big Data*.

https://goo.gl/RD3Tis

Referências

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *Artificial Intelligence Magazine*, v. 17, n. 3, p. 37-54, 1996. Disponível em: <https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>. Acesso em: 23 dez. 2018.

GONÇALVES, E. C. *Extração de Árvores de Decisão com a Ferramenta de Data Mining Weka*. 2007. Disponível em: <https://www.devmedia.com.br/images/articles/168773/arvore.jpg>. Acesso em: 23 dez. 2018.

HEKIMA. *Se você se interessa por Big Data, precisa entender o CRISP-DM*. 2016. Disponível em: <http://www.bigdatabusiness.com.br/se-voce-se-interessa-por-big-data-precisa-entender-o-crisp-dm/>. Acesso em: 23 dez. 2018.

LANHELLAS, R. *Redes Neurais Artificiais*: Algoritmo Backpropagation 2013. Disponível em: <https://arquivo.devmedia.com.br/artigos/Ronaldo_Lanhellas/Algoritmo-Backpropagation/Algoritmo-Backpropagation2.jpg>. Acesso em: 23 dez. 2018.

SALVADOR, H. G.; CUNHA, A. M.; CORRÊA, C. S. Vedalogic: um método de verificação de dados climatológicos apoiado em modelos minerados. *Revista Brasileira de Meteorologia*, v. 24, n. 4, p. 448-460, dez. 2009. Disponível em: <http://www.scielo.br/scielo.php?pid=S0102-77862009000400007&script=sci_abstract&tlng=es>. Acesso em: 23 dez. 2018.

Leituras recomendadas

BIG DATA BUSINESS. *Aplicações de Big Data*. 2018. Disponível em: <http://www.bigdata-business.com.br/aplicacoes-de-big-data/>. Acesso em: 23 dez. 2018.

BRITO, M. *Aspectos teóricos da mineração de dados e aplicação das regras de classificação para apoiar o comércio*. 2012. Disponível em: <https://www.devmedia.com.br/aspectos-teoricos-da-mineracao-de-dados-e-aplicacao-das-regras-de-classificacao-para-apoiar-o-comercio/25429>. Acesso em: 23 dez. 2018.

CAMILO, C. O.; SILVA, J. C. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Relatório técnico*, p. 1–29, ago. 2009. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 23 dez. 2018.

DUNCAN, O. et al. *Conceitos de mineração de dados*. 2018. Disponível em: <https://docs.microsoft.com/pt-br/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017>. Acesso em: 23 dez. 2018.

FERREIRA, R. S. *10 ferramentas e bibliotecas para trabalhar com data mining e Big Data*: Parte 01. 2017. Disponível em: <https://imasters.com.br/data/10-ferramentas-e-bibliotecas-para-trabalhar-com-data-mining-e-big-data-parte-01>. Acesso em: 23 dez. 2018.

HEKIMA. *Big Data*: tudo que você sempre quis saber sobre o tema! 2016. Disponível em: <http://www.bigdatabusiness.com.br/tudo-sobre-big-data/>. Acesso em: 23 dez. 2018.

PIATETSKY, G. R. *Python Duel As Top Analytics, Data Science software*: KDnuggets 2016 Software Poll Results. 2016. Disponível em: <https://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html>. Acesso em: 23 dez. 2018.

RAMAKRISHNAN, R.; GEHRKE, J. *Sistemas de gerenciamento de banco de dados*. 3. ed. Porto Alegre: Penso, 2013.

Conteúdo: