INTRODUÇÃO A BIG DATA E INTERNET DAS COISAS (IOT)

Priscila Gonçalves

Utilizar técnicas de *Data Mining*

Objetivos de aprendizagem

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

- Identificar as técnicas de *Data Mining*.
- Reconhecer a lógica para *Data Mining*.
- Aplicar a sintaxe de consultas de *Data Mining*.

Introdução

O processo de *Data Mining* — mineração de dados — assume cada vez mais relevância nos mais diversos contextos, já que, com cada vez mais frequência, lidamos com uma quantidade imensa de dados que precisam ser explorados, analisados e organizados. Por isso, existem diversas técnicas que auxiliam os profissionais nesse processo, fornecendo informações sobre os dados, e há um interesse crescente pelo desenvolvimento de novos métodos.

Neste capítulo, você estudará as principais técnicas de mineração de dados, verá como reconhecer a lógica para *Data Mining* e aprenderá a aplicar a sintaxe de consultas de mineração.

Identificação das técnicas de *Data Mining*

Variados tipos de dados podem ser minerados e, para tal, podem ser utilizadas técnicas diferentes. O processo para a criação de um modelo de mineração representa uma parte de um processo maior, que inclui perguntas sobre dados e no qual consta, inclusive, um modelo de respostas para as perguntas feitas e a implantação do modelo propriamente dito.

Os métodos ou técnicas de mineração de dados podem ser divididos em supervisionado (preditivo) e não supervisionado (descritivo), por esforço e semissupervisionado. A diferença entre o supervisionado e o não supervisionado se

dá pelo fato de que os não supervisionados não necessitam de pré-categorização para registros, de modo que não se faz necessário ter um atributo-alvo — necessitamos de menos informações sobre os objetos, segundo Daniil Korbut (2017).

Dentre os algoritmos que podem ser utilizados nesse processo, pode-se citar: associação, classificação, *clustering*, árvores de decisão e padrões sequenciais.

Mineração por grupo de associação

O método de mineração por grupo de associação tem como propósito identificar elementos que tenham a presença de outros elementos em uma mesma operação, encontrando relacionamentos ou padrões entre o conjunto de dados — a transação mostra os itens que foram consultados em uma determinada operação. Essas regras de associação representam padrões em transações armazenadas e, por meio do conhecimento desses dados, organizações podem direcionar processos de marketing e promover estratégias que tragam vantagens. Geralmente, as bases de dados envolvidas nesses tipos de processos são muito grandes e, para elas, é necessária a utilização de algoritmos rápidos e eficientes.

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

Regra 1: SE idade = jovem AND trabalha = não ENTÃO compra notebook = não

Regra 2: SE idade = jovem AND trabalha = sim ENTÃO compra notebook = sim

Regra 3: SE idade = adulto AND crédito = sim ENTÃO compra carro = sim

Regra 4: SE idade = adulto AND crédito = não ENTÃO compra carro = não

Mineração por classificação

Nesta técnica, vários atributos podem ser utilizados para a identificação de uma classe específica de itens. São atribuídos itens às categorias ou classes de destino pela classificação, de forma que possa ser previsto com uma maior precisão o que poderá ocorrer dentro das classes. Essa é uma técnica que, geralmente, é utilizada dentro do marketing para classificar o público para suas campanhas.

A Figura 1, a seguir, representa a mineração de dados por classificação, relacionando idade e classe social.

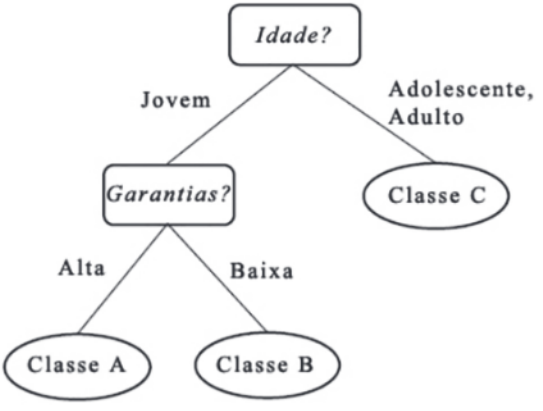

Figura 1. Mineração de dados por classificação.
Fonte: Camilo e Silva (2009, p. 13).

Mineração por *clustering*

A técnica de *clustering* agrupa registros semelhantes, ou seja, grupos de elementos que possuem as mesmas propriedades a fim de que o usuário final possa, entre outras coisas, saber o que está ocorrendo no banco de dados. Essa técnica é bastante utilizada pelo marketing para saber quais objetos podem ajudar na segmentação, como, por exemplo, segmentando o mercado em subconjuntos de clientes, em que cada um desses subconjuntos poderá ser direcionado para uma estratégia de marketing diferente, com padrões diferentes para diferentes tipos de clientes.

A Figura 2 representa a mineração de dados por *clustering*.

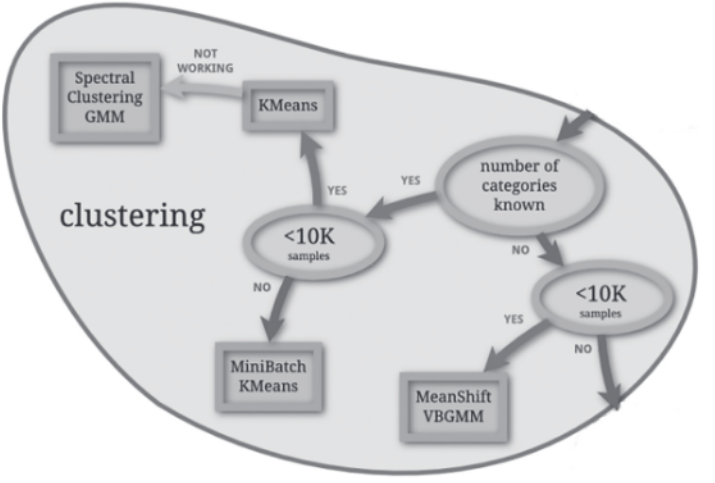

Figura 2. Mineração de dados por *clustering*.
Fonte: Analytics Vidhya (2018, documento on-line).

Mineração por árvores de decisão

A técnica de mineração por árvores de decisão é utilizada para categorização ou previsão de dados. Geralmente, inicia com uma pergunta (caracterizada por um conjunto de dados de entrada) que tenha duas ou mais respostas (dados de saídas); cada uma dessas respostas direciona para uma questão que será utilizada para classificar ou identificar dados que serão categorizados ou, ainda, poderá ser feita uma previsão baseada em cada resposta. Ou seja, uma árvore de decisão é formada a partir de um conjunto de regras de classificação, e cada caminho da raiz até uma folha representa uma dessas regras. Normalmente, a árvore de decisão é definida de forma que, para cada observação pertencente à base de dados, haja um e somente um caminho da raiz até a folha.

A Figura 3, a seguir, representa uma árvore de decisão para comprar, ou não, algo, com base em preço e nível de importância de consumo.

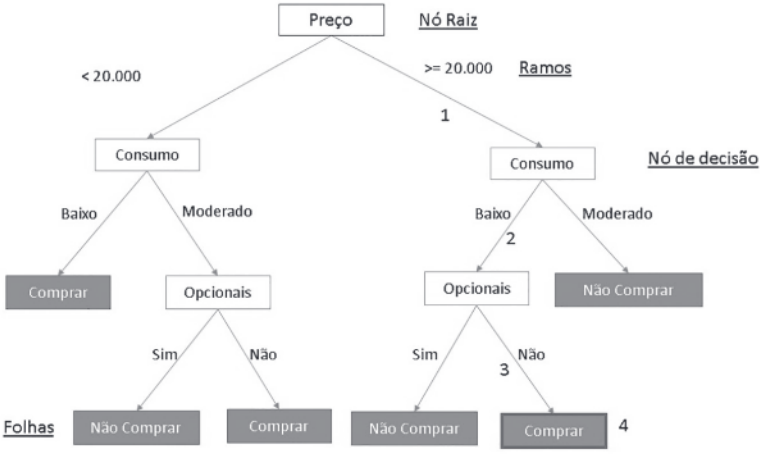

Figura 3. Mineração de dados por árvore de decisão.

Fonte: Devmedia (2014, documento on-line).

Mineração por padrões sequenciais

Padrões, geralmente, identificam tendências ou a ocorrência de eventos parecidos. Essa técnica de mineração de dados costuma ser utilizada para entender comportamentos de usuários em relação às compras, em que os donos de lojas, a partir da análise dos dados, tomam decisões sobre quais produtos irão apresentar para os clientes. Um exemplo em que pode ser utilizada essa técnica é em "carrinhos de compras" de lojas on-line, pois, a partir de um histórico de compras do cliente, o empreendedor pode sugerir que algo mais seja adicionado às compras.

Fique atento

Pode-se pensar nas diferentes tarefas de mineração de dados como consultas complexas, com especificação em alto nível, com parâmetros definidos pelo usuário e os algoritmos especializados que serão implementados a elas.

A lógica para a mineração de dados

Por meio da mineração de dados, é possível descobrir informações de grande valor, principalmente para ajudar nas tomadas de decisões. Para a realização da mineração de dados, são aplicados algoritmos específicos, nos quais são especificadas regras e aplicadas lógicas para que as informações obtidas sejam as desejadas pelo usuário. Após a aplicação dos algoritmos em um conjunto de dados, os resultados são sintetizados e são aplicadas ferramentas que apoiarão a decisão de mineração.

Algoritmos de mineração de dados são conjuntos de heurística e cálculos que criam modelos com base nos dados. Para que ocorra a criação do modelo, o algoritmo realiza a análise de dados que são fornecidos a ele e, a partir disso, ocorre a busca por tipos de padrões ou tendências específicas. Dessa forma, o algoritmo utiliza resultados dessa análise em diversas iterações, definindo parâmetros ideais para a criação do modelo de mineração. Esses parâmetros deverão ser aplicados pelo conjunto de dados para extrair padrões acionáveis e estatísticas com mais riqueza de detalhes.

O modelo de mineração criado pelo algoritmo pode assumir vários formatos, como um conjunto de *clusters*, uma árvore de decisão, modelos matemáticos ou um conjunto de regras que descreverá como serão agrupados produtos em uma transação e as probabilidades de que os produtos sejam comprados juntos. Dentre os métodos mais populares e bem conhecidos para derivar padrões de dados, estão (AYODELE, 2010):

- Classificação (supervisionada) — prevê que cada registro faça parte de outro conjunto de dados e pertença a uma determinada classe.
- Regressão (supervisionada) — prevê que cada registro faça parte de outro conjunto de dados e tenha um determinado valor.
- Estimação (supervisionada) — envolve a geração de pontuação para cada registro.
- Clusterização (não supervisionada) — identifica grupos que poderão ser utilizados como ponto inicial de exploração de relação, procurando semelhanças e diferenças em conjunto de dados e agrupando registros semelhantes em segmentos ou *clusters*.
- Associação (não supervisionada) — gera modelos descritivos que proporcionam o descobrimento de regras.

- Análise de sequenciação (não supervisionada) — interessa a ordem em que aparecem nas transações e o espaço de tempo entre elas.
- Visualização — apresentação gráfica dos dados.

Aplicação e sintaxe de consultas de mineração

Consulta de conteúdo é uma maneira de extrair informações sobre as estatísticas internas e a estrutura do modelo de mineração. Uma consulta de conteúdo pode fornecer detalhes que estarão disponíveis de maneira acessível no visualizador, e os resultados podem ser utilizados para extrair informações para outras utilizações. Essas consultas podem retornar padrões, fórmulas, lista de atributos e todas as informações que forem julgadas como pertinentes ao negócio. Consultas sobre estrutura e dados armazenados em *cache* são, geralmente, utilizados para criar estruturas de mineração e modelos.

Diante da intensificação de pesquisas na área de desenvolvimento de algoritmos, pode-se contar com uma grande oferta de ferramentas para a mineração de dados, tanto gratuitas quanto pagas. Entre elas, pode-se citar Weka, Mahout, Orange Data Mining, Rapid Miner, Tanagra, Keel. Em relação às alternativas de ferramentas pagas, pode-se citar Oracle, Microsoft, SAS entre outras. E, mesmo diante dessa gama de ferramentas, um dos principais desafios consiste em saber identificar qual estratégia melhor se aplica ao contexto, questão e problema que buscam ser solucionados. Dentre as linguagens de programação mais utilizadas para mineração de dados, pode-se citar Python, principalmente pela simplicidade, clareza e reusabilidade. Trata-se de uma linguagem de sintaxe simples e objetiva que permite aos programadores manter o foco no problema a ser resolvido sem que haja preocupações com implementações. R também é uma linguagem de programação poderosa quando o tema é *data science* (ciência de dados), pois tem facilidade para analisar dados, processar instruções estatísticas e modelos gráficos.

A Figura 4 apresenta as ferramentas e linguagens de programação mais utilizadas na mineração de dados com base em uma pesquisa realizada em 2016, na 17ª edição anual do KDnuggets Software Poll.

Tool	2016 % share	% change	% alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

Figura 4. Ferramentas e linguagens de programação mais utilizadas em *Data Mining* com base em pesquisa em 2016.

Fonte: Piatetsky (2016, documento on-line).

Para obter todos os dados incluídos na estrutura, assim como as colunas que não foram adicionadas a um modelo de mineração específico, você deverá ter permissões de detalhamento no modelo, assim como na estrutura, para recuperar dados da estrutura de mineração.

A partir da consulta de conteúdo modelo, é possível:

- extrair fórmulas ou probabilidades para fazer seus próprios cálculos;
- em um modelo de associação, recuperar as regras que são usadas para gerar uma previsão;
- recuperar as descrições de regras específicas para usá-las em um aplicativo personalizado;
- apresentar as médias móveis detectadas por um modelo de série temporal;
- obter a fórmula de regressão para algum segmento da linha de tendência;
- recuperar informações acionáveis sobre clientes identificados como fazendo parte de um *cluster* específico.

Exemplo

Exemplos de utilização de **SQL** para *data mining*:
Exemplo para obter intervalo de valores, localizando o valor mínimo e o valor máximo:
SELECT DISTINCT RangeMin(<column>), RangeMax(<column>) FROM <model>

Exemplo para recuperar informações detalhadas sobre nós específicos no modelo. De acordo com o tipo de algoritmo, o nó pode conter regras e fórmulas, suporte e estatísticas de variância, e assim por diante:
SELECT FROM <model>.CONTENT

Exemplo para recuperar conteúdo armazenado em uma dimensão de mineração de dados:
SELECT FROM <model>.DIMENSIONCONTENT

Essa instrução retorna tipos diferentes de informações dependendo do modelo que está sendo consultado. Em um exemplo de mineração de dados por associação, uma informação importante é o tipo de nó. Segue um exemplo para a consulta em um **modelo de associação**:

```
SELECT FLATTENED NODE_UNIQUE_NAME, NODE_DESCRIPTION,
   (SELECT RIGHT(ATTRIBUTE_NAME, (LEN(ATTRIBUTE_NAME)-LEN('Association model
name')))
FROM NODE_DISTRIBUTION
WHERE LEN(ATTRIBUTE_NAME)>2
)
AS RightSideProduct
FROM [<Association model name>].CONTENT
WHERE NODE_TYPE = 8
ORDER BY NODE_SUPPORT DESC
```

Exemplo para consulta em um modelo de **árvore de decisão** que pode ser utilizado tanto para previsão quanto para classificação:

```
SELECT  Predict([Bike Buyer]), PredictNodeID([Bike Buyer])
FROM [<decision tree model name>]
PREDICTION JOIN
<input rowset>
```

Os nós são usados para representar árvores e nós folha; dessa forma, para rastrear o caminho para qualquer resultado específico, será necessário identificar o nó que o contém e obter os detalhes para aquele nó.

Veja, a seguir, um exemplo de utilização da **linguagem R**, que apresenta a sintaxe básica da língua e a utilização da interface da linha de comando. Para apresentar os resultados de uma função, a biblioteca dplyr usa o operador (%>%).
No exemplo a seguir, serão apresentados apenas filmes de suspense que tenham duração maior do que 120 minutos e que tenham sido lançados em 2015. Serão mostrados como resultados os títulos e a classificação:

```
library(dplyr)
  movies %>%
  filter(year == 2015 & thriller == 1 & length > 115) %>%
  select(title, rating)
```

Saiba mais

Acesse o link a seguir e saiba mais a respeito das consultas de mineração e suas aplicações.

https://goo.gl/Szeu92

Referências

ANALYTICS VIDHYA. *Drop shadows background*. 2018. Disponível em: <https://www.analyticsvidhya.com/blog/2017/02/top-28-cheat-sheets-for-machine-learning-data-science-probability-sql-big-data/drop_shadows_background/>. Acesso em: 24 dez. 2018.

AYODELE, T. O. Types of machine learning algorithms. In: ZHANG, Y. (Ed.). *New advances in machine learning*. London: IntechOpen, 2010. Disponível em: <http://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms>. Acesso em: 24 dez. 2018.

CAMILO, C. O.; SILVA, J. C. *Mineração de Dados:* conceitos, tarefas, métodos e ferramentas. 2009. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 24 dez. 2018.

DEVMEDIA. *Mineração de dados com árvores de decisão.* 2014. Disponível em: <https://www.devmedia.com.br/mineracao-de-dados-com-arvores-de-decisao/31397>. Acesso em: 24 dez. 2018.

KORBUT, D. *Machine Learning Algorithms:* Which One to Choose for Your Problem. 26 out. 2017. Disponível, em: <https://blog.statsbot.co/machine-learning-algorithms-183cc73197c>. Acesso em: 24 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.

Leituras recomendadas

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.

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.

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

EDSON. *Trabalhando com a linguagem R*. 2015. Disponível em: <https://www.devmedia.com.br/trabalhando-com-a-linguagem-r/33275>. Acesso em: 23 dez. 2018.

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.

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.

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

SILVA, J. C. *Algoritmos de Aprendizagem de Máquina:* qual deles escolher? 2018. Disponível em: <https://medium.com/machina-sapiens/algoritmos-de-aprendizagem-de-m%C3%A1quina-qual-deles-escolher-67040ad68737>. Acesso em: 24 dez. 2018.

WU, X. et al. Dez algoritmos em mineração de dados. *Knowledge and Information Systems*, v. 14, p. 1-37, 2008. Disponível em: <https://mineracaodedados.files.wordpress.com/2012/05/ten-algorithms-in-data-mining.pdf>. Acesso em: 23 dez. 2018.

Conteúdo: