

Leandro Augusto da Silva

# Mineração de dados

Uma abordagem  
introdutória e  
ilustrada



Editora  
**Mackenzie**

# **Mineração de dados**

Uma abordagem  
introdutória e  
ilustrada



11

**UNIVERSIDADE PRESBITERIANA MACKENZIE**

*Reitor: Benedito Guimarães Aguiar Neto*

*Vice-reitor: Marcel Mendes*

**EDITORA DA UNIVERSIDADE PRESBITERIANA MACKENZIE**

**Conselho Editorial**

Helena Bonito Couto Pereira (*Presidente*)

José Francisco Siqueira Neto

Leila Figueiredo de Miranda

Luciano Silva

Maria Cristina Triguero Veloz Teixeira

Maria Lucia Marcondes Carvalho Vasconcelos

Moises Ari Zilber

Valter Luís Caldana Júnior

Wilson do Amaral Filho

**COLEÇÃO CONEXÃO INICIAL**

*Diretora: Maria Lucia Marcondes Carvalho Vasconcelos*

Leandro Augusto da Silva

# Mineração de dados

Uma abordagem  
introdutória e  
ilustrada

 Editora  
**Mackenzie**

© 2015 Leandro Augusto da Silva

Todos os direitos reservados à Editora Mackenzie.  
Nenhuma parte desta publicação poderá ser reproduzida por qualquer meio  
ou forma sem a prévia autorização da Editora Mackenzie.

Coordenação editorial: Joana Figueiredo  
Produtora editorial: Andréia Ferreira Cominetti

Capa: Rubens Lima  
Preparação de texto: Eugênia Pessotti  
Diagramação: Sidnei Simonelli/Printfit  
Revisão: Claudia Silveira e Hebe Lucas

**Dados Internacionais de Catalogação na Publicação (CIP)**  
**(Câmara Brasileira do Livro, SP, Brasil)**

---

Silva, Leandro Augusto da

Mineração de dados : uma abordagem introdutória e ilustrada / Leandro Augusto da Silva. -- 1. ed. -- São Paulo : Editora Mackenzie, 2015. -- (Coleção conexão inicial ; v. 11)

Bibliografia

ISBN: 978-85-8293-059-5

1. Banco de dados - Pesquisas
2. Mineração de dados (Computação)
3. Processamento de dados I. Título. II. Série.

14-10846

CDD-005.741

---

Índice para catálogo sistemático:

1. Mineração de dados : Banco de dados : Computadores :  
Processamento de dados 005.741

EDITORA MACKENZIE  
Rua da Consolação, 930  
Edifício João Calvino  
São Paulo – SP – CEP 01302-907  
Tel.: (5511) 2114-8774 (editorial)  
editora@mackenzie.br  
www.mackenzie.br/editora.html

Como adquirir o livro:  
Cia. dos Livros – Mackenzie  
Rua da Consolação, 930  
Edifício João Calvino – térreo  
São Paulo – SP – CEP 01302 907  
Tel.: (5511) 3129-4319  
mackenzie@ciadoslivros.com.br  
www.ciadoslivros.com.br/

# SUMÁRIO

<b>Sobre o autor</b>	<b>7</b>
<b>Lista de figuras</b>	<b>9</b>
<b>Lista de tabelas</b>	<b>11</b>
<b>Apresentação</b>	<b>13</b>
<b>1 Introdução</b>	<b>17</b>
1.1 Outras abordagens para análise de dados	18
1.2 Descoberta de conhecimento em banco de dados	26
1.2.1 Base de dados	28
1.2.2 Pré-processamento	32
Preparação	32
Seleção	36
Transformação	39
1.2.3 Mineração de dados	41
1.2.4 Análise dos resultados	42
1.3 Exemplo de contexto	42
1.4 Exercícios de aplicação	46
<b>2 Classificação de dados</b>	<b>51</b>
2.1 Árvore de decisão ou DT ( <i>decision tree</i> )	54
2.1.1 Exemplo de contexto para DT	57
2.2 <i>K</i> -vizinhos mais próximos	70
2.2.1 Exemplo de contexto para <i>k</i> -NN	75
2.3 Avaliação de modelos	79
2.4 Exercícios de aplicação	82

<b>3</b>	<b>Clusterização de dados</b>	<b>89</b>
3.1	Agrupamento hierárquico	92
3.1.1	Exemplo de contexto para agrupamento hierárquico aglomerativo	98
3.2	$k$ -médias	105
3.2.1	Exemplo de contexto para $k$ -médias	108
3.3	Validação de agrupamento	113
3.4	Exercícios de aplicação	115
<b>4</b>	<b>Associação de dados</b>	<b>119</b>
4.1	Apriori	121
4.1.1	Exemplo de contexto para o algoritmo Apriori	128
4.2	FP-Growth	132
4.1.2	Exemplo de contexto para o algoritmo FP-Growth	135
4.3	Exercícios de aplicação	139
	<b>Referências</b>	<b>143</b>
	<b>Bibliografia comentada</b>	<b>147</b>
	<b>Glossário</b>	<b>149</b>
	<b>Índice</b>	<b>151</b>

# SOBRE O AUTOR

## Leandro Augusto da Silva

Graduado em Engenharia da Computação pela Faculdade de Engenharia Industrial (FEI), Leandro Augusto da Silva é mestre e doutor em Engenharia Elétrica pela Escola Politécnica da Universidade de São Paulo. É professor na Universidade Presbiteriana Mackenzie, vinculado à Faculdade de Computação e Informática (FCI) e ao programa de pós-graduação *strictu-sensu* de Engenharia Elétrica. Leciona disciplinas relacionadas a processamento de imagens, banco de dados, redes neurais artificiais, *big data* e mineração de dados. Na mesma instituição, participa de comissões de ensino, núcleos temáticos e coordena atividades complementares. É revisor técnico de conferências nacionais e internacionais na área de inteligência computacional e de revistas especializadas. Publica regularmente artigos científicos nos principais congressos nacionais e internacionais de sua área de pesquisa, bem como em revistas da área. É um dos organizadores do livro *Tendências tecnológicas em computação e informática*, também lançado pela Editora Mackenzie.

# LISTA DE FIGURAS

Figura 1.1	Exemplo de três relações de banco de dados: funcionário, cargo e formação	19
Figura 1.2	Cubo <i>warehouse</i> para exploração de dimensões de um repositório analítico de dados	23
Figura 1.3	Resultado de um BI para o exemplo da base de dados de contexto	25
Figura 1.4	Processo iterativo e interativo para descoberta de conhecimento em bases de dados	27
Figura 1.5	Etapas do pré-processamento	28
Figura 1.6	Taxonomia do tipo de atributos de uma base de dados	31
Figura 1.7	Resultados de correlação de Pearson para nota 1 com nota 2 (a) e horas de estudo com nota 2 (b)	39
Figura 1.8	Tarefas de mineração de dados	41
Figura 2.1	Representação genérica para treinamento e teste de um modelo preditivo	52
Figura 2.2	Exemplo genérico de uma árvore de decisão e sua interpretação	56
Figura 2.3	Construção do ramo à esquerda da DT	59
Figura 2.4	Árvore de decisão resultante do processo de treinamento a partir da base de dados de conhecimento	59
Figura 2.5	Exemplares divididos por cada valor do atributo EP	63
Figura 2.6	Divisão dos exemplares com os atributos EP e QR combinados	66
Figura 2.7	Separação de ramos para mais de dois valores possíveis de um atributo	68
Figura 2.8	Divisão de ramos para atributos numéricos	69
Figura 2.9	Plano cartesiano com exemplares distribuídos no espaço de atributos	71
Figura 2.10	Exemplo de classificação com o $k$ -NN	75

Figura 2.11	Gráfico de dispersão com os dados da base <i>ANÚNCIOS</i>	<b>77</b>
Figura 2.12	Representação pictórica do método de validação <i>Leave-one-out</i> .	<b>80</b>
Figura 3.1	Exemplares descritos por atributos regulares, sem atributo especial classe e sem formação objetiva do número de grupos	<b>90</b>
Figura 3.2	Diferentes bases de dados com exemplares distribuídos sobre diferentes estruturas de grupo	<b>91</b>
Figura 3.3	Processo para construção de um dendrograma com cinco exemplares	<b>94</b>
Figura 3.4	Exemplo de agrupamento hierárquico e resultado com dendrograma, com critério <i>single-link</i> (ligação mais próxima)	<b>98</b>
Figura 3.5	Exemplo de cortes no dendrograma para descoberta <i>k</i> grupos	<b>98</b>
Figura 3.6	Gráfico de dispersão com os dados da base <i>CARTÃO DE CRÉDITO</i>	<b>101</b>
Figura 3.7	Iterações do algoritmo aglomerativo para a construção do dendrograma	<b>102</b>
Figura 3.8	Exemplo pictórico de agrupamento de dados com <i>k</i> -médias	<b>107</b>
Figura 3.9	Gráfico de dispersão para a base de dados, notas de alunos e ilustração dos centroides iniciais e atualizados	<b>113</b>
Figura 4.1	Diagrama de Venn para o <i>itemset</i> {A,D}	<b>125</b>
Figura 4.2	Caminho atualizado para $t_1$	<b>134</b>
Figura 4.3	Caminho atualizado para $t_2$	<b>134</b>
Figura 4.4	Caminho atualizado após $t_3, t_4, t_5$ e $t_6$	<b>134</b>
Figura 4.5	Execução do algoritmo FP-Growth com a base <i>SUPERMERCADO</i>	<b>138</b>

# LISTA DE TABELAS

Tabela 1.1	Armazéns de dados ( <i>data warehouse</i> ) para processamento analítico	22
Tabela 1.2	Armazéns de dados ( <i>data warehouse</i> ) com uso função agregada <i>count</i>	22
Tabela 1.3	Base de dados de candidatas a uma vaga de emprego	29
Tabela 1.4	Exemplo de inconsistência de dados	35
Tabela 1.5	Base de dados de exemplos para notas de alunos. Os valores das notas variam de 0 a 10, os valores das horas de estudos variam em minutos e a situação pode ser A – aprovado e R – reprovado	38
Tabela 1.6	Atributos regulares da base alunos normalizados com a técnica <i>min-max</i>	40
Tabela 1.7	Base de dados com pesquisas de restaurantes	43
Tabela 1.8	Resumo da base de dados de restaurantes	43
Tabela 1.9	Base de conhecimento	46
Tabela 1.10	Base de dados com <i>ANÚNCIOS</i> em páginas da internet	49
Tabela 2.1	Base restaurante pré-processada para construção de uma DT	58
Tabela 2.2	Matriz de contadores para organizar os exemplares do atributo <i>EP</i>	63
Tabela 2.3	Resultado de Impureza para os atributos QR e LE da Tabela 2.1	65
Tabela 2.4	Matriz de contagens para o atributo EP e ramo Pouca com o atributo QR	65
Tabela 2.5	Base de dados <i>ANÚNCIOS</i> pré-processada para classificação com <i>k</i> -NN	76
Tabela 2.6	Resultado da comparação (distância) entre o exemplar $x_1$ e os demais da base de dados	78
Tabela 2.7	Resultado das comparações organizado em ordem crescente	78
Tabela 2.8	Resultado para atribuição de classe	78
Tabela 2.9	Representação da matriz de confusão para duas classes	82

Tabela 3.1	Base de dados <i>CARTÃO DE CRÉDITO</i>	<b>100</b>
Tabela 3.2	Matriz de similaridade para a base de dados <i>CARTÃO DE CRÉDITO</i>	<b>101</b>
Tabela 3.3	Método alternativo para visualização da formação de grupos pelo algoritmo aglomerativo	<b>104</b>
Tabela 3.4	Base de dados com <i>NOTAS DOS ALUNOS</i>	<b>109</b>
Tabela 3.5	Resultado da comparação de cada exemplar com os centroides e associação dos exemplares para a primeira iteração do algoritmo <i>k</i> -médias	<b>110</b>
Tabela 3.6	Recálculo do centroide $c_1$ para o grupo	<b>111</b>
Tabela 3.7	Recálculo do centroide $c_2$ para o grupo	<b>111</b>
Tabela 3.8	Resultado da comparação de cada exemplar com os centroides e associação dos exemplares para a segunda iteração do algoritmo <i>k</i> -médias	<b>112</b>
Tabela 3.9	Teste de estabilidade na iteração 1 e 2	<b>112</b>
Tabela 3.10	Base de dados <i>NOTAS DOS ALUNOS</i> com adição do atributo especial de classe, o qual refere-se ao grupo descoberto pelo <i>k</i> -médias	<b>113</b>
Tabela 4.1	Exemplo de base transacional usada para associação de dados	<b>121</b>
Tabela 4.2	Resultado do suporte para a primeira iteração do Algoritmo 4.1	<b>124</b>
Tabela 4.3	Resultado do suporte para a combinação de itens	<b>125</b>
Tabela 4.4	Base de dados <i>SUPERMERCADO</i> usada como exemplo de contexto para descoberta de regras de associação	<b>129</b>
Tabela 4.5	Resultado da primeiro passo do Algoritmo 4.1 para o exemplo de contexto	<b>129</b>
Tabela 4.6	Resultado de suporte para a combinação de dois itens ( $L_2$ )	<b>130</b>
Tabela 4.7	Resultado de suporte para a combinação de três itens ( $L_3$ )	<b>131</b>
Tabela 4.8	Apresentação do cálculo da confiança para o conjunto de itens resultante	<b>131</b>
Tabela 4.9	Itens ordenados pelo suporte para exemplo ilustrativo	<b>133</b>
Tabela 4.10	Produtos com suporte maior que o mínimo e ordenados pelo suporte para a base <i>SUPERMERCADO</i>	<b>135</b>

# APRESENTAÇÃO

A imensa quantidade de dados gerados atualmente tem feito com que a capacidade humana para analisá-los e interpretá-los seja extrapolada. O armazenamento digital de dados, que há pouco era objeto de desejo de grandes e médias empresas, agora se torna um desafio para pesquisadores e corporações, no que diz respeito a manipular analiticamente essa superabundância de dados. A esse desafio soma-se o interesse em determinar ações estratégicas, visando à descoberta de conhecimento em bases de dados para aumentar vendas, definir perfis e sugerir produtos relacionados. A descoberta de conhecimento constitui um processo, cuja primeira etapa tem o objetivo de fazer um pré-processamento na base de dados para entregar os dados limpos, preparados e selecionados à fase seguinte. Nesta fase, que é a principal, está a mineração de dados, na qual algoritmos de aprendizado de máquina ou de redes neurais artificiais são executados sobre os dados, a fim de criar um modelo que auxilie em tarefas como modelagem preditiva (regressão e classificação), análise de clustering ou agrupamento e regras de associação. Como última etapa, os resultados da mineração são interpretados e analisados, qualitativamente e quantitativamente.

Diante do exposto, nota-se que essa é uma área interdisciplinar e exige do leitor uma grande diversidade de experiências que envolvem basicamente banco de dados, álgebra linear, matemática discreta e algoritmos. Nesse sentido, esta obra tem como objetivo a apresentação destes assuntos, de forma contextualizada, de modo a facilitar o entendimento de um problema e sua resolução por meio de algoritmos escritos em pseudocódigos e executados passo a passo. Com esta estratégia, o livro constitui-se de uma visão bastante pragmáti-

ca dos algoritmos de mineração de dados e suas aplicações em casos reais.

Quanto ao desenvolvimento dos capítulos, todos seguirão estrutura similar, facilitando o entendimento dos conceitos. Cada capítulo tem início com uma introdução, na qual apresentamos a técnica proposta, os algoritmos disponíveis na literatura e as aplicações. O intuito é oferecer uma visão geral ao leitor e caminhos, na forma de referências, para que possa se aprofundar no tema. Nessa introdução, ainda apresentamos os algoritmos que serão estudados. O critério adotado para estas escolhas foi a opção pelos dois mais populares de cada tarefa de mineração de dados. Como corpo principal do capítulo, os algoritmos são apresentados em detalhes, porém em profundidades diferentes. No início introduz-se a ideia, a partir de exemplo pictórico. Conseqüentemente, apresentamos o algoritmo formal, em pseudocódigo, escrito em um português estruturado, e executamos passo a passo, por meio de um exemplo simples. Por fim, como último nível de profundidade, apresentamos um problema real para contextualizar a tarefa de mineração de dados que está sendo discutida, e apresentamos a solução a esse problema com o uso do algoritmo estudado no capítulo. Os capítulos são finalizados com uma conclusão por meio de destaques aos pontos principais dos algoritmos e, também, dos pontos falhos e de como esses são resolvidos por outras técnicas da literatura. Para a obtenção de uma melhor ideia de cada capítulo, seguem a estrutura do livro, os objetivos e os assuntos abordados:

## **Capítulo 1 - Introdução**

Apresenta ao leitor o processo usado para descobrir conhecimento em base de dados. As etapas desse processo serão amplamente discutidas por meio de exemplos teóricos e práticos, com exceção da etapa “mineração de dados”, na qual apenas introdu-

ziremos suas principais tarefas: modelagem preditiva, análise de clustering e regras de associação de dados.

Ao final do capítulo, espera-se que o leitor compreenda:

- O que é mineração de dados;
- A diferença entre mineração de dados, sistemas de processamento transacional, sistemas de processamento analítico e inteligência de negócios;
- O processo de descoberta de conhecimento em bases de dados.

## Capítulo 2 - Classificação de dados

Trata do conceito de classificação de dados. Veremos, neste capítulo, dois algoritmos tradicionais de classificação: árvores de decisão (*decision tree*) e  $k$  vizinhos mais próximos (*k nearest neighbor*). Os algoritmos serão amplamente discutidos e contextualizados em exemplos reais de mineração de dados.

Ao final desse capítulo, espera-se que o leitor compreenda:

- A diferença entre classificação e previsão da modelagem preditiva;
- O algoritmo árvores de decisão;
- O algoritmo  $k$ -NN;
- A construção de um processo de classificação;
- As técnicas para avaliação de resultados de classificação.

## Capítulo 3 - Clusterização de dados

Aborda o conceito da tarefa clusterização ou agrupamento de dados. Estudaremos, neste capítulo, dois importantes algoritmos, agrupamento hierárquico e  $k$ -médias. Assim como no capítulo anterior, discutiremos amplamente estes dois algoritmos com exemplos contextualizados.

Ao final desse capítulo, espera-se que o leitor compreenda:

- O que é agrupamento de dados e sua importância;
- O algoritmo agrupamento hierárquico;
- O algoritmo  $k$ -médias;
- A construção de um processo de agrupamento de dados;
- As medidas de validação de agrupamento.

## Capítulo 4 - Associação de dados

Abordaremos, neste capítulo, a tarefa de mineração de dados conhecida como associação de dados. Essa tarefa é realizada em um processo de duas fases, sendo o primeiro para encontrar itens frequentes e o seguinte para gerar regras de associação. Para o primeiro processo, temos dois principais algoritmos, chamados Apriori e FP-Growth. Discutiremos, no capítulo, ambos os algoritmos, a partir de exemplos contextualizados, e ainda faremos aplicações práticas.

Ao final desse capítulo, espera-se que o leitor compreenda:

- A importância de gerar regras de associação em um processo de descoberta de conhecimento;
- O algoritmo Apriori;
- O algoritmo FP-Growth;
- A construção de um processo de associação de dados.

*Mineração de dados: uma abordagem introdutória e ilustrada* apresenta esse importante conceito da computação usado para análise de dados nas mais variadas áreas, como indústria, agronegócio, medicina, *marketing* e segurança da informação. Este livro traz uma visão pragmática da mineração de dados, com modelos em três níveis de assimilação: exemplificação, contextualização e aplicação. Com linguagem objetiva e didática, a obra explica algoritmos de forma prática, fornecendo ao estudante os conhecimentos necessários à implementação das tarefas de mineração de dados e à indicação de soluções para problemas reais.

