CONEXÃO INICIAL

Leandro Augusto da Silva

Mineração de dados

Uma abordagem introdutória e ilustrada

Editora Mackenzie

Mineração de dados

Uma abordagem introdutória e ilustrada

CONEXÃO INICIAL

11

Leandro Augusto da Silva

Mineração de dados

Uma abordagem introdutória e ilustrada

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

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

LISTA DE TABELAS

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. Consequentemente, 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.

ISBN: 978-85-8293-059-5

CONEXÃO INICIAL