



Data Mining

2ª
edição

*Conceitos, técnicas,
algoritmos, orientações
e aplicações*





Ronaldo Goldschmidt
Emmanuel Passos
Eduardo Bezerra

Data Mining

2ª
edição

*Conceitos, técnicas,
algoritmos, orientações
e aplicações*



Material
na WEB
www.elsevier.com.br


CAMPUS

© 2015, Elsevier Editora Ltda.

Todos os direitos reservados e protegidos pela Lei nº 9.610, de 19/02/1998.

Nenhuma parte deste livro, sem autorização prévia por escrito da editora, poderá ser reproduzida ou transmitida, sejam quais forem os meios empregados: eletrônicos, mecânicos, fotográficos, gravação ou quaisquer outros.

Revisão: Marco Antonio Corrêa

Editoração Eletrônica: WM Design

Elsevier Editora Ltda.

Conhecimento sem Fronteiras

Rua Sete de Setembro, 111 – 16º andar

20050-006 – Centro – Rio de Janeiro – RJ

Rua Quintana, 753 – 8º andar

04569-011 – Brooklin – São Paulo – SP

Serviço de Atendimento ao Cliente

0800 026 53 40

atendimento1@elsevier.com

ISBN 978-85-352-7822-4

ISBN (versão eletrônica) 978-85-352-7823-1

NOTA

Muito zelo e técnica foram empregados na edição desta obra. No entanto, podem ocorrer erros de digitação, impressão ou dúvida conceitual. Em qualquer das hipóteses, solicitamos a comunicação ao nosso serviço de Atendimento ao Cliente para que possamos esclarecer ou encaminhar a questão.

Nem a editora nem o autor assumem qualquer responsabilidade por eventuais danos ou perdas a pessoas ou bens, originados do uso desta publicação.

CIP-Brasil. Catalogação na publicação.
Sindicato Nacional dos Editores de Livros, RJ

G575d

2. ed.

Goldschmidt, Ronaldo

Data mining : conceitos, técnicas, algoritmos, orientações e aplicações / Ronaldo

Goldschmidt, Eduardo Bezerra. - 2. ed. - Rio de Janeiro : Elsevier, 2015.

il. ; 24 cm.

Inclui bibliografia

ISBN 978-85-352-7822-4

1. Mineração de dados (Computação). 2. Banco de dados. I. Bezerra, Eduardo. II.

15-21831

CDD: 005.74

CDU: 004.65

13/04/2015 20/04/2015

Prefácio

Conforme o próprio título sugere, este livro reúne desde material teórico e formal até experiências e orientações práticas reais sobre como conduzir e executar aplicações na área da descoberta de conhecimento em bases de dados (KDD). Seu conteúdo é uma introdução aos conceitos fundamentais necessários para se realizar o processo de KDD.

Durante todo o livro, exemplos são utilizados para demonstrar a aplicação das técnicas de KDD em situações práticas. Ao fim de cada capítulo, são fornecidos exercícios para que o leitor possa verificar o entendimento do conteúdo apresentado.

Público-alvo

Este livro é destinado a estudantes de nível técnico, de graduação e pós-graduação em Informática, Computação ou em Engenharia que estejam cursando alguma disciplina introdutória sobre KDD. Profissionais de outras áreas como a Estatística também podem encontrar neste livro uma boa iniciação aos conceitos computacionais de KDD e à sua aplicação.

Em geral, o livro é adequado para profissionais de Tecnologia da Informação interessados em utilizar dados históricos para extrair conhecimento que possa ser utilizado na tomada de decisões. Assim, o texto mescla uma abordagem conceitual e formal com linguagem acessível, recomendada a todos os tipos de leitores, seguido de informações de cunho mais prático, voltado ao público com interesse na aplicação da tecnologia.

O conhecimento dos fundamentos de programação e de banco de dados é desejável (mas não obrigatório) para o bom entendimento dos assuntos tratados.

Organização dos capítulos

O Capítulo 1 apresenta uma visão geral da área de KDD e sua relação com outras áreas do conhecimento. Contém uma retrospectiva histórica sobre o tema e a interpretação dos autores acerca da diversidade de expressões relacionadas

cada vez mais populares como Mineração de Dados, Big Data, Ciência de Dados, dentre outras.

O Capítulo 2 complementa a visão geral da área de KDD e Mineração de Dados, com conceitos básicos necessários aos capítulos subsequentes. Também apresenta os tipos de profissionais que atuam na área.

O Capítulo 3 enfoca a Descoberta de Conhecimento em Bases de Dados como processo, detalhando suas etapas e algumas das funções de KDD mais utilizadas. Um mesmo exemplo de banco de dados é considerado ao longo do capítulo de modo que o leitor possa acompanhar de forma encadeada todo o processo.

Sumarização, Classificação, Agrupamento, Descoberta de Associações e Previsão de Séries Temporais estão entre as principais tarefas de KDD apresentadas no Capítulo 4.

O Capítulo 5 mostra diversos métodos e técnicas de Mineração de Dados. Para um melhor aproveitamento do conteúdo apresentado, recomenda-se (para os leitores sem conhecimento em técnicas de Inteligência Computacional) que a leitura desse capítulo seja precedida por um estudo do material complementar (veja o item Recursos da Web deste prefácio) que contém noções introdutórias sobre Redes Neurais, Lógica Nebulosa e Algoritmos Genéticos.

O Capítulo 6 trata de um tópico de grande importância no processo de Descoberta de Conhecimento em Bases de Dados: a caracterização de uma metodologia para orientar o processo de KDD. Baseada na CRISP-DM (modelo industrial que contém diretrizes para execução de aplicações de KDD), tal metodologia e seus mecanismos de controle encontram-se descritos em detalhe nesse capítulo.

Com a popularização de técnicas para a análise de redes sociais on-line e dos sistemas de recomendação, muita atenção tem sido dedicada à análise de dados representados como grafos. Grafos são estruturas matemáticas amplamente utilizadas na representação abstrata de dados. O Capítulo 7 dedica-se à apresentação de tarefas e aplicações de análises de dados estruturados em grafos, como, por exemplo, predição de ligações e a detecção de comunidades.

Bastante popular na atualidade, a expressão Big Data se refere ao conjunto de técnicas e procedimentos que abrangem a coleta contínua, a integração, o armazenamento e a análise dinâmica de dados, possivelmente esparsos, provenientes de várias fontes e em diferentes formatos. O Capítulo 8 apresenta uma introdução ao cenário de Big Data, destacando conceitos e tecnologias de NoSQL e de Mineração de Dados Paralela e Distribuída (com introduções ao MapReduce e ao Hadoop).

Nos Capítulos 9 e 10 são apresentadas, a título ilustrativo, algumas das principais experiências práticas reais vivenciadas pelos autores em projetos envolvendo Mineração de Dados. Em particular, o Capítulo 10 destaca aplicações de Mineração de Dados Educacionais, enfatizando os resultados obtidos com

o apoio do projeto Memore e do programa USA-UCA no contexto “Um Computador por Aluno”.

Por fim, o Capítulo 11 resume as principais tendências nas áreas de KDD e de Mineração de Dados, além de fornecer algumas orientações para os leitores que atuam ou pretendam atuar na área.

Recursos na web

Como informação suplementar à contida neste livro, disponibilizamos um site na própria editora Elsevier/Campus. O leitor pode acessar a página da editora (www.elsevier.com.br). Nesse endereço, o leitor pode obter informações e material complementar sobre o tema. O leitor pode também utilizar esse site para entrar em contato com os autores, com o objetivo de trocar ideias sobre o livro. Entre os recursos que podem ser encontrados no site, estão os seguintes:

- Soluções de alguns dos exercícios propostos no livro.
- Apresentações baseadas no conteúdo dos assuntos abordados no livro. Esse material é útil para o professor ou instrutor que deseja adotar o livro em seus cursos.
- Implementações de alguns dos métodos e/ou técnicas apresentados no livro.
- Um texto introdutório sobre Redes Neurais Artificiais, Lógica Nebulosa, Algoritmos Genéticos e Data Warehouse. Recomenda-se que leitores não familiarizados com os temas leiam esse texto antes do Capítulo 5 do livro.
- Outras fontes de informação. O material disponível no site da editora contém também endereços para outras fontes de informação sobre KDD. Seguindo a natureza dinâmica da internet, o conteúdo do site será modificado de tempos em tempos.

Agradecimentos

Registramos nossos agradecimentos às pessoas que, de alguma forma, colaboraram para a elaboração desta segunda edição.

Ao Fabio de Azevedo, pelas valiosas ideias de melhoria em diversos pontos do texto e pela ajuda com muitas das figuras e das referências utilizadas.

Aos professores Maria Claudia Cavalcanti, Claudia Justel, Julio Duarte e Ricardo Choren, docentes do Instituto Militar de Engenharia (IME) e membros da equipe do projeto PredLig, pelo aprimoramento do conteúdo e do texto do Capítulo 7. Em particular, ao Ricardo Choren também pela revisão do texto do Capítulo 11.

A toda a equipe do projeto Memore e do programa USA-UCA e da Secretaria Municipal de Educação de Pirai (RJ), pelas oportunidades, ações e resultados obtidos com a mineração dos dados educacionais, experiência relatada no Capítulo 9.

Aos professores Daniel Oliveira (Instituto de Computação – UFF) e Glauco Amorim (CEFET/RJ), pelo apoio na elaboração e revisão dos textos sobre as arquiteturas de hardware e computação em nuvem apresentados no Capítulo 8.

Aos discentes Ana Paula Teixeira, Carlos Henrique Moreira, Elaine da Costa Tady, Gustavo Costa, Jessica Aparecida Seibert, Jones Marques, Josiane Oliveira e Juliane Marinho, bolsistas de iniciação científica do curso de Ciência da Computação da UFRRJ, pelo apoio na elaboração dos manuais de utilização do Weka e do Rapid Miner.

Aos pesquisadores e professores Artur Ziviani (LNCC), Fabio Porto (LNCC), Jonice Oliveira (PPGI – UFRJ) e Sergio Serra (UFRRJ) por gentilmente ceder alguns de seus trabalhos que serviram de fontes para elaboração de partes deste livro.

Ao Professor Eduardo Ogasawara (CEFET/RJ), por valiosas contribuições e discussões acerca do conteúdo do Capítulo 11.

Finalmente, mas não menos importante, agradecemos às nossas famílias por todo apoio e incentivo, fundamentais para que pudéssemos finalizar esta obra.

Convite ao leitor

Convidamos o leitor a prosseguir pelo restante desta obra. Esperamos que as informações nela contidas sejam úteis e que a leitura seja a mais agradável possível. Nossos votos são de que o conteúdo introdutório apresentado neste livro possa despertar o interesse do leitor pela área de KDD e Mineração de Dados, e, de alguma forma, contribuir para sua formação profissional.

Tentamos ao máximo produzir um texto cuja leitura seja aprazível e didática. Entretanto, pelo fato de a produção de um livro ser uma tarefa bastante complexa, temos consciência de que erros e inconsistências ainda se escondem por entre as linhas que compõem esta edição. Para os que quiserem entrar em contato conosco para trocar ideias e fornecer críticas e sugestões, fiquem à vontade para enviar mensagens.

Ronaldo Goldschmidt, Eduardo Bezerra e Emmanuel Passos
Rio de Janeiro, janeiro de 2015

Sumário

Prefácio	v
-----------------------	----------

CAPÍTULO 1

Introdução	1
1.1 KDD: uma visão geral.....	1
1.2 Dado, informação, conhecimento	2
1.3 Definições de KDD.....	4
1.3.1 <i>Perspectiva do conhecimento extraído</i>	4
1.3.2 <i>Perspectiva da realização do processo</i>	9
1.4 Áreas relacionadas com a KDD	10
1.4.1 <i>Aprendizado de máquina</i>	10
1.4.2 <i>Estatística</i>	11
1.4.3 <i>Bancos de Dados e Data Warehousing</i>	12
1.5 Atividades de KDD	12
1.6 Perspectiva histórica da área de KDD	13
Exercícios	16

CAPÍTULO 2

Processo de KDD: Conceitos Básicos	19
2.1 Caracterização do Processo de KDD.....	19
2.1.1 <i>Definição do problema</i>	19
2.1.2 <i>Recursos disponíveis</i>	20
2.1.3 <i>Resultados obtidos</i>	21
2.2 Etapas operacionais do processo de KDD	22
2.2.1 <i>Pré-processamento</i>	23
2.2.2 <i>Mineração de dados</i>	24
2.2.3 <i>Pós-processamento</i>	27
2.3 Macroobjetivos e orientações do processo de KDD	27
2.4 Operações de KDD	28

2.5 Métodos de KDD.....	29
2.6 Técnicas de KDD.....	29
2.7 Ferramentas de KDD.....	32
2.8 O papel do usuário no processo de KDD.....	32
Exercícios	34

CAPÍTULO 3

Etapas do Processo de KDD	35
3.1 Considerações iniciais.....	35
3.2 Pré-processamento.....	38
3.2.1 Seleção de dados	38
3.2.2 Limpeza	52
3.2.3 Codificação	55
3.2.4 Enriquecimento dos dados	60
3.2.5 Normalização de dados	61
3.2.6 Construção de atributos	66
3.2.7 Correção de prevalência.....	66
3.2.8 Partição do conjunto de dados	67
3.3 Mineração de Dados	69
3.3.1 Representação do conhecimento	69
3.3.2 Medidas de interesse.....	70
3.3.3 Similaridade e distância	71
3.3.4 Aprendizado indutivo	73
3.4 Pós-processamento.....	74
3.4.1 Simplificações do modelo de conhecimento	75
3.4.2 Transformações de modelo de conhecimento	76
3.4.3 Organização e apresentação dos resultados	77
Exercícios	78

CAPÍTULO 4

Tarefas de KDD	81
4.1 Considerações iniciais.....	81
4.2 Tarefas Primárias	81
4.2.1 Descoberta de regras de associação	81
4.2.2 Descoberta de Associações Generalizadas.....	84
4.2.3 Descoberta de Sequências	85
4.2.4 Descoberta de Sequências Generalizadas.....	88
4.2.5 Classificação.....	88
4.2.6 Regressão	92
4.2.7 Sumarização	95

4.2.8 Clusterização/Agrupamento.....	95
4.2.9 Previsão de Séries Temporais	101
4.2.10 Detecção de Desvios.....	103
4.3 Tarefas compostas.....	103
4.3.1 Clusterização → Classificação	103
4.3.2 Clusterização → Sumarização.....	104
4.4 Meta-Aprendizado.....	104
Exercícios	109

CAPÍTULO 5

Métodos de Mineração de Dados.....	115
5.1 Considerações iniciais.....	115
5.2 Métodos Tradicionais.....	117
5.2.1 Comentários gerais	117
5.2.2 <i>k</i> -NN.....	117
5.2.3 Classificador Bayesiano Ingênuo	121
5.2.4 <i>k</i> -Means.....	125
5.2.5 <i>Apriori</i>	128
5.2.6 <i>C4.5</i>	131
5.2.7 Máquinas de Vetores Suporte.....	137
5.3 Métodos Bioinspirados.....	145
5.3.1 Métodos baseados em Redes Neurais	145
5.3.2 Métodos baseados em Algoritmos Genéticos.....	150
5.3.3 Métodos baseados em Lógica Nebulosa	160
5.4 Tarefas de KDD e Métodos de Mineração de Dados: Resumo	163
Exercícios	109

CAPÍTULO 6

Metodologia de KDD.....	167
6.1 Considerações iniciais.....	167
6.2 CRISP-DM.....	168
6.2.1 Compreensão do Negócio	169
6.2.2 Compreensão dos Dados.....	169
6.2.3 Preparação dos Dados.....	170
6.2.4 Modelagem.....	171
6.2.5 Avaliação	171
6.2.6 Desenvolvimento.....	171
6.3 Metodologia Proposta.....	171
6.3.1 Levantamento da Situação Vigente.....	174
6.3.2 Definição de Objetivos	174

6.3.3 Planejamento de atividades	175
6.3.4 Execução dos Planos de Ação	177
6.3.5 Avaliação de Resultados	178
6.4 Considerações complementares	179
Exercícios	178

CAPÍTULO 7

Mineração de Grafos	183
7.1 Considerações iniciais	183
7.2 Introdução aos grafos	184
7.2.1 Terminologia	184
7.2.2 Métricas	188
7.2.3 Representações computacionais	189
7.3 Tarefas	191
7.3.1 Predição de Ligações	191
7.3.2 Detecção de comunidades	194
7.3.3 Ranqueamento	196
7.4 Aplicações	199
7.4.1 Web Mining	199
7.4.2 Sistemas de recomendação baseados em grafos	202
7.4.3 Análise de Redes Sociais	206
Exercícios	209

CAPÍTULO 8

Big Data	211
8.1 Considerações iniciais	211
8.2 Fundamentos e tecnologias relacionadas	212
8.2.1 Arquiteturas de hardware	212
8.2.2 Mineração de Dados Paralela vs. Distribuída	214
8.2.3 Tipos de Paralelismo	215
8.2.4 Estratégias de Balanceamento de Cargas de Trabalho	216
8.2.5 Estratégias de distribuição de Modelos de Conhecimento	216
8.2.6 Decomposição de bases de dados	217
8.2.7 Computação em Nuvem	217
8.2.8 MapReduce	221
8.3 Tarefas	223
8.3.1 Mineração de Regras de Associação e Apriori Paralelo	224
8.3.2 Metaclassificação distribuída	224

8.4 Tecnologias Relacionadas.....	225
8.4.1 Hadoop®	225
8.4.2 NoSQL	226
Exercícios	232

CAPÍTULO 9

Mineração de Dados Educacionais	233
9.1 Considerações iniciais	233
9.2 Um breve histórico	234
9.3 Uma taxonomia para aplicações da EDM	235
9.4 Exemplos de Aplicação da EDM	236
9.4.1 O Projeto Memore	236
9.4.2 O programa USA-UCA	247
9.4.3 Outras aplicações em EDM	254
9.5 Considerações complementares	254
Exercícios	256

CAPÍTULO 10

Exemplos de Aplicação de KDD	257
10.1 Considerações iniciais.....	257
10.2 Telefonia	257
10.3 Franquia de fast-food	258
10.4 Ação social	258
10.5 Área Médica.....	258
10.6 Área Financeira	259
10.7 Outros exemplos	259
Exercícios	261

CAPÍTULO 11

Considerações finais	263
11.1 Retrospectiva	263
11.2 Tendências e perspectivas	264

Referências	269
--------------------------	------------

Listagem de figuras

Figura 1.1. Hierarquia entre Dado, Informação e Conhecimento	2
Figura 1.2. Perspectiva gráfica de um conjunto de dados de clientes de uma instituição financeira. Adaptado de Naliato (2001).....	8
Figura 1.3. Taxonomia de atividades na área de KDD.....	13
Figura 2.1. Etapas Operacionais do Processo de KDD.....	22
Figura 2.2. Modelo Híbrido Sequencial	31
Figura 2.3. Modelo Híbrido Auxiliar	31
Figura 2.4. Modelo Híbrido Incorporado	32
Figura 2.5. Ser humano como elemento central do processo de KDD.....	33
Figura 3.1. Visualização de um conjunto de dados de três dimensões numéricas.....	43
Figura 3.2. Conjunto de dados resultante da projeção	44
Figura 3.3. Abordagens para seleção de atributos: (a) Wrapper; (b) Filter.....	46
Figura 3.4. Exemplo de procedimento para arredondamento de valores.....	50
Figura 3.5. Resultados dos Métodos de Redução de Valores.....	51
Figura 3.6. Exemplo de Enriquecimento de Dados.....	60
Figura 3.7. Hipóteses de funções induzidas a partir dos exemplos de entradas e saídas	74
Figura 3.8. Exemplo de Árvore de Decisão e suas regras.....	77
Figura 4.1. Associações entre registros de dados e classes.....	89
Figura 4.2. Conhecimento extraído a partir dos dados na Tabela 4.6	92
Figura 4.3. Estágios da Metaclassificação	106
Figura 4.4. (a) Estratégia de arbitragem; (b) estratégia de combinação.....	107
Figura 5.1. Conjunto contendo dados sobre clientes que receberam crédito.....	118
Figura 5.2. Seleção da vizinhança do registro “*” durante o processamento do k-NN no exemplo apresentado ($k = 3$).....	119
Figura 5.3. Comparação entre os comportamentos das estratégias de interpolação com pesos uniformes <i>versus</i> com pesos decrescentes com a distância	121
Figura 5.4. Diagrama de atividade do algoritmo k-Means.....	126
Figura 5.5. Primeiros passos do algoritmo k-Means.....	127
Figura 5.6. Passos subsequentes do algoritmo k-Means.....	127

Figura 5.7. Em um espaço de dimensão 2, um hiperplano separador é uma reta	137
Figura 5.8. Diferentes retas separam o conjunto de dados	138
Figura 5.9. Vetores de suporte determinam o Classificador Linear de margem máxima	139
Figura 5.10. Erros de classificação cometidos pelo Classificador Linear	141
Figura 5.11. Mapeamento de um conjunto de dados de R1 para R2. Adaptada de (HAMEL, 2009, p. 95)	143
Figura 5.12. Ilustração gráfica de um Mapa de Kohonen	150
Figura 5.13. Ilustração da redução de vizinhança em um Mapa de Kohonen	150
Figura 5.14. Representação de cromossoma – Rule Evolver	153
Figura 5.15. Crossover Lógico do Rule Evolver	154
Figura 5.16. Mutação Lógica do Rule Evolver	154
Figura 5.17. Representações de cromossomas para clusters (x_1, x_3, x_6) , (x_2, x_4, x_5) : (a) grupamento de número; (b) matriz; (c) permutação com o caractere 7 representado como o separador dos dois clusters; (d) permutação greedy	155
Figura 5.18. Representação do crossover de cromossomas redundantes gerando cromossomas inválidos	156
Figura 5.19. Crossover assexual de 1 ponto	157
Figura 5.20. Crossover sexual de 2 pontos	157
Figura 5.21. Operadores de crossover para restauração de cromossoma por permutação. (a) Crossover baseado em mapeamento parcial; (b) crossover baseado em ordem	158
Figura 5.22. Operador de mutação para representação do cromossoma por matriz	159
Figura 5.23. Exemplo de série temporal dividida em 7 conjuntos nebulosos	161
Figura 6.1. Fases do Modelo de Referência CRISP-DM. Adaptado de Shearer (2000)	168
Figura 6.2. Formulário para Documentação de Ações e Resultados do Processo de KDD	173
Figura 6.3. Exemplo de preenchimento do formulário na etapa de Definição de Objetivos	175
Figura 6.4. Alternativas de Planos de Ação	176
Figura 6.5. Exemplo de preenchimento do formulário na etapa Planejamento de Atividades	177
Figura 6.6. Exemplo de preenchimento do formulário na etapa Execução de Planos de Ação	178
Figura 7.1. Exemplo de digrafo simples com 4 vértices e 3 arestas	185
Figura 7.2. Exemplos de grafos completos	185
Figura 7.3. Exemplo de grafo bipartido	186
Figura 7.4. Exemplo de grafo com laço	186
Figura 7.5. Grafo conexo e com uma ponte	187
Figura 7.6. Exemplos de representações computacionais para o grafo K_3	190
Figura 7.7. Exemplo de grafo com três comunidades	194
Figura 7.8. Funcionamento do algoritmo de Girvan-Newman	195
Figura 7.9. Dendrograma, estrutura produzida pelo algoritmo Girvan-Newman	196
Figura 7.10. Grafo para ilustrar o funcionamento do PageRank	198
Figura 7.11. Três abordagens de Web Mining	200

Figura 8.1. Máquinas de Memória Distribuída (DMM).....	213
Figura 8.2. Máquina de Memória Compartilhada (SMM - SMP).....	213
Figura 8.3. Clusters de SMPs – CLUMPs. Fonte: Zaki, 2000.....	214
Figura 8.4. Taxonomia para computação em nuvem adaptada de Oliveira (2010).....	219
Figura 8.5. Exemplo de estrutura tipicamente utilizada em Bigtables	229
Figura 8.6. Fragmento de documento com dados de clientes de uma empresa	231
Figura 9.1. Abstração da organização dos dados do Memore – visão simplificada	239
Figura 9.2. Interface entre agente de coleta de dados e: (a) aluno alfabetizado ; (b) aluno em alfabetização.....	241
Figura 9.3. Interface dos questionários sobre: (a) a situação operacional da escola; (b) o perfil socioeconômico docente.....	241
Figura 9.4. Relatório gerencial sobre a distribuição do tempo de uso dos laptops por disciplina.....	242
Figura 9.5. Tempo médio de uso dos laptops – Distribuição por: (a) disciplina; (b) local de uso	244
Figura 9.6. Evolução do desempenho acadêmico médio das turmas-piloto em Matemática	245
Figura 9.7. Evolução do tempo de utilização dos laptops por turma-piloto	245
Figura 9.8. Exemplos de características do perfil socioeconômico docente	246
Figura 9.9. Perfil de utilização dos laptops UCA na E.M. Rosa Carelli da Costa.....	251
Figura 9.10. Perfil de utilização dos laptops UCA no Ciep 477 Professora Rosa Guedes.....	253

Listagem de tabelas

Tabela 1.1. Exemplo de padrão preditivo	9
Tabela 3.1. Estrutura do conjunto de dados de clientes.....	35
Tabela 3.2. Exemplos de instâncias de clientes.....	36
Tabela 3.3. Classificação das variáveis do conjunto de dados de clientes	37
Tabela 3.4. Autovetores e autovetores correspondentes da matriz de covariância	44
Tabela 3.5. Resultado do agrupamento dos atributos sexo e estado civil da Tabela 3.2	48
Tabela 3.6. Intervalos com comprimento definido pelo usuário.....	57
Tabela 3.7. Intervalos divididos com igual comprimento.....	57
Tabela 3.8. Divisão em intervalos em função do tamanho da amostra	57
Tabela 3.9. Exemplo de Representação Binária Padrão (Econômica).....	58
Tabela 3.10. Exemplo de Representação Binária 1-de-N	59
Tabela 3.11. Representação Binária por Temperatura	59
Tabela 3.12. Distância de Hamming entre os conceitos da Tabela 3.11.....	59
Tabela 3.13. Exemplo de Normalização Linear.....	62
Tabela 3.14. Exemplo de Normalização por Desvio Padrão.....	63
Tabela 3.15. Normalização pela soma dos elementos.....	64
Tabela 3.16. Exemplo de normalização pelo valor máximo dos elementos	64
Tabela 3.17. Exemplo de normalização por escala decimal.....	65
Tabela 3.18. Medidas de similaridade para dados categóricos	72
Tabela 3.19. Exemplo de conjunto de dados	76
Tabela 4.1. Relação das vendas de um minimercado em um período.....	82
Tabela 4.2. Formato Cesta da relação das vendas da Tabela 4.1.....	83
Tabela 4.3. Relação das compras realizadas por cliente.....	86
Tabela 4.4. Exemplos de sequências do conjunto de dados da Tabela 4.3	87
Tabela 4.5. Matriz de Confusão de um Classificador – problema com k classes.....	91
Tabela 4.6. Matriz de Confusão de um Classificador – problema com 2 classes.....	91
Tabela 4.7. Clientes e suas compras em um tipo de literatura.....	92
Tabela 4.8. Dados dos funcionários de uma empresa fictícia	94
Tabela 5.1. Back-Propagation e C4.5 representados por precondições e efeitos.....	116
Tabela 5.2. Exemplos de Precondições e Efeitos de Métodos de Pré-Processamento.....	116

Tabela 5.3. Conjunto de dados para ilustrar o comportamento do CBI.....	124
Tabela 5.4. Conjunto das vendas de um minimercado fictício.....	129
Tabela 5.5. Conjunto de dados para ilustrar a indução de Árvores de Decisão.....	134
Tabela 5.6. Formulação do problema de otimização do SVM.....	140
Tabela 5.7. Formulação do problema de otimização por meio de Multiplicadores de Lagrange.....	140
Tabela 5.8. Formulação do problema de otimização considerando ruídos nos dados.....	142
Tabela 5.9. Mapeamento entre os valores do atributo tipo de residência e os símbolos respectivamente representados.....	153
Tabela 5.10. Exemplos de Tarefas de KDD e Métodos de Mineração de Dados	163
Tabela 6.1. Execução de Processos de KDD – Metodologia Proposta x CRISP-DM	173
Tabela 7.1. Índices de centralidade (grau, proximidade e intermediação)	189
Tabela 7.2. Passos do algoritmo Girvan-Newman.....	195
Tabela 7.3. Valores de diversas iterações do PageRank.....	198
Tabela 8.1. Exemplo de conjunto de pares {chave, valor} produzido pela função Map	222
Tabela 8.2. Exemplo de conjunto de pares {chave, valor} produzido pela função Reduce	223
Tabela 8.3. Exemplos de aplicação do padrão chave-valor	228
Tabela 9.1. Grupos de interesse na EDM.....	235
Tabela 9.2. Classificação de tarefas da EDM segundo Romero & Ventura (2010).....	237
Tabela 9.3. Distribuição dos laptops e tempos de utilização pelas turmas-piloto.....	243
Tabela 9.4. Principais aspectos operacionais levantados nas escolas-piloto (7/2012).....	245
Tabela 9.5. Regras de Associação geradas pelo Apriori (SupMin = 3%; ConfMin = 70%).....	247
Tabela 9.6. Regras de Associação geradas pelo Apriori (SupMin = 2%; ConfMin = 80%).....	252
Tabela 9.7. Regras de associação geradas pelo Apriori (SupMin = 10%; ConfMin = 70%).....	252