A Mineração de Dados, ou *data mining* ganhou evidência após o surgimento do termo *Big Data*, em que a mineração de dados é o elemento central responsável pela preparação e análise das grandes massas de dados.

Com a nova nomenclatura, até os profissionais que atuam na área ganharam novo nome: *cientistas de dados*. E esses profissionais são cada vez mais requisitados, principalmente no momento em que o volume de dados produzidos cresce exponencialmente, ao ponto de que em curtos períodos se gera mais dados do que em muitos séculos de história da humanidade.

Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações traz uma abordagem introdutória e sistemática ao tema, fornecendo, para cada tarefa da mineração de dados, uma descrição do processo, conceitos básicos, exemplos ilustrados, pseudocódigos para os principais algoritmos, além de exercícios de fixação disponibilizados *on-line*.

Indicação:

A obra é indica para as disciplinas de Mineração de Dados (ou Data Mining); Aprendizagem de Máquina (ou Machine Learning); Descoberta de Conhecimento em Bases de Dados (ou Knowledge Discovery in Databases – KDD); Análise de Dados; Big Data; Processos Analíticos; Inteligência Analítica; Data Warehouse; Business Intelligence. É também de extrema valia para profissionais que atuam nas áreas de Mineração de Dados, Big Data, Modelagem e Análise de dados, Segurança, Inteligência Etratégica, Inteligência de Marketing e Processos Analíticos.

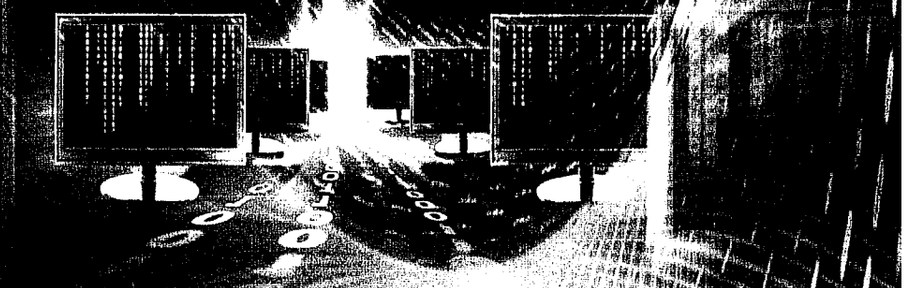
ISBN 978-85-472-0098-5

LEANDRO NUNES DE CASTRO
DANIEL GOMES FERRARI
INTRODUÇÃO À MINERAÇÃO DE DADOS

Editora Saraiva

LEANDRO NUNES DE CASTRO
DANIEL GOMES FERRARI

INTRODUÇÃO À MINERAÇÃO DE DADOS



CONCEITOS BÁSICOS, ALGORITMOS E APLICAÇÕES

Editora Saraiva

Editora Saraiva

Rua Henrique Schaumann, 270
Pinheiros – São Paulo – SP – CEP: 05413-010
PABX (11) 3613-3000

SAC | **0800-0117875**
De 2ª a 6ª, das 8h30 às 19h30
www.editorasaraiva.com.br/contato

Diretora editorial	Flávia Alves Bravin
Gerente editorial	Rogério Eduardo Alves
Planejamento editorial	Rita de Cássia S. Puoço
Editores	Ana Laura Valerio do Nascimento Fernando Alves Fernando Penteado Isabella Sanches Patricia Quero
Assistente editorial	Marcela Prada Neublum
Produtores editoriais	Alline Garcia Bullara Amanda Maria da Silva Daniela Nogueira Secondo Deborah Mattos Rosana Peroni Fazolari William Rezende Paiva
Comunicação e produção digital	Mauricio Scervianinas de França Nathalia Setrini Luiz
Suporte editorial	Juliana Bojczuk
Produção gráfica	Liliane Cristina Gomes
Preparação	Jean Xavier
Revisão	Rosângela Barbosa
Diagramação	Join Bureau
Capa	Guilherme P. Pinto
Impressão e acabamento	Corprint Gráfica e Editora Ltda.

ISBN 978-85-472-0098-5

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
ANGÉLICA ILACQUA CRB-8/7057

de Castro, Leandro Nunes
Introdução à mineração de dados : conceitos básicos, algoritmos e aplicações / Leandro Nunes de Castro, Daniel Gomes Ferrari. – São Paulo : Saraiva, 2016.

Bibliografia
ISBN 978-85-472-0098-5

1. Mineração de dados (Computação) I. Título II. Ferrari, Daniel Gomes

16-0023 CDD 006.312
CDU 004.89

Índices para catálogo sistemático:
1. Mineração de dados (Computação)

1ª edição

383.472.001.001

Porque a diferença entre dado, informação e conhecimento é maior que nossa capacidade de transformar um no outro de maneira rápida e precisa, principalmente para massas de dados com grandes volumes, variedades e velocidades.

1

Introdução à mineração de dados

As coisas estão mudando [...] Na última década, a maior parte do trabalho que fazemos [...] migrou para o computador, agora ligado a uma rede. Estamos ligados a um colega equipado com uma memória fenomenal, um senso estranho de tempo e nenhuma lealdade [...] corremos o risco de nos tornarmos servos e escravos da informação que produzimos.

BAKER, S. *The Numeratti*, 2009, p. 18.

NESTE CAPÍTULO, VOCÊ ESTUDARÁ:

- O que é mineração de dados, quais as principais tarefas e receberá dicas para uma análise eficiente e eficaz
- O significado das diferentes nomenclaturas, como Inteligência Artificial e Aprendizagem de Máquina, dentre outras
- Exemplos de aplicação

1.1 INTRODUÇÃO

A quantidade de usuários da internet no mundo todo passou de 16 milhões de pessoas em 1995 para aproximadamente 2,8 bilhões em 2013 (Figura 1.1(a)); a quantidade de artigos publicados apenas em inglês na Wikipédia passou de 500 mil em 2005 para quase 4,4 milhões em 2013; o tempo necessário para o rádio atingir uma audiência de 50 milhões de pessoas foi de 38 anos, ao passo que a TV precisou de 13 anos e a internet, de apenas quatro anos para alcançar esse mesmo número de pessoas;[1] a quantidade de buscas diárias no Google ultrapassa cinco bilhões, são escritos 500 milhões de tuítes por dia e vistas 200 milhões de horas de vídeos no YouTube diariamente. Ainda no YouTube, foram enviadas 13 milhões de horas de vídeo apenas no ano 2010, o que corresponde a aproximadamente oito anos de conteúdo enviados todos os dias.[2]

Figura 1.1 Tempos exponenciais

(a) Crescimento do número de usuários de internet no mundo (Fonte: Internetworldstats.com).
(b) Lei de Moore: crescimento do número de transistores em um circuito integrado (Fonte: Intel.com).

[1] TAPSCOTT, D. *Grown up digital*. Nova York: McGraw-Hill, 2009.

[2] De acordo com estatísticas divulgadas pelo próprio YouTube, disponíveis em: <youtube.com/t/press_statistics>. Acesso em: 28 jan. 2016.

Em 1965, Gordon Moore, um dos fundadores da Intel, publicou um artigo[3] no qual observou que a quantidade de componentes em um circuito integrado (CI) estava dobrando aproximadamente a cada ano desde a sua invenção, em 1958, e essa taxa permaneceria por pelo menos mais dez anos. Em 1975, Moore atualizou sua estimativa para períodos de dois anos, em vez de um ano. Essa elevada taxa de crescimento na quantidade de componentes do CI está diretamente relacionada à velocidade de processamento e capacidade de memória dos computadores e também tem servido de meta para a indústria de hardware computacional (Figura 1.1(b)).

Paradoxalmente, esses avanços da tecnologia – tanto de hardware quanto de comunicação – têm produzido um problema de superabundância de dados, pois a capacidade de coletar e armazenar dados tem superado a habilidade de analisar e extrair conhecimento destes. Nesse contexto, é necessária a aplicação de técnicas e ferramentas que transformem, de maneira inteligente e automática, os dados disponíveis em informações úteis, que representem conhecimento para uma tomada de decisão estratégica nos negócios e até no dia a dia de cada um de nós. Nesse sentido, pesquisadores das mais variadas áreas têm se dedicado a estudar métodos para *análise de dados*.

Ao observar as duas curvas da Figura 1.1(a), percebe-se um comportamento similar entre elas – mais acentuado no caso da Figura 1.1(b) –, que é um aumento significativamente maior da função no eixo vertical em relação ao aumento no eixo horizontal, chamado de *crescimento exponencial*. O crescimento de usuários de internet, de artigos publicados na rede, de tuítes diários e de vídeos assistidos no YouTube, tudo segue um crescimento exponencial – é exatamente esse o momento em que vivemos: *tempos exponenciais*. Em tempos exponenciais, encontrar e acessar fontes de informação, pessoas, produtos e serviços não é mais problema; na verdade, o desafio atual é gerenciar, armazenar, processar e extrair conhecimento a partir dessa quantidade quase ilimitada de dados.

Uma das mais emblemáticas representantes dessa superabundância de dados é a *computação em nuvem* (*cloud computing*), que se refere ao fornecimento de recursos computacionais como serviço em vez de produto. Na nuvem, recursos (hardware e software) são compartilhados por meio de uma rede, geralmente a internet, e os usuários podem acessar as aplicações hospedadas em servidores remotos utilizando browsers, desktops e até aplicativos móveis. Essa possibilidade tem atraído muitas empresas e usuários comuns, principalmente em virtude dos elevados custos e da complexidade de manutenção de servidores de dados e aplicação.

Para termos uma ideia da dimensão que a nuvem vem ganhando, em abril de 2012 o Greenpeace publicou um relatório intitulado "Quão limpa é sua nuvem?",[4] no qual é feita uma análise da quantidade e do tipo de fonte de energia consumida por cada uma das maiores empresas de computação do planeta, como Amazon, Apple, Facebook, Google, Microsoft, Twitter e Yahoo!. De acordo com o relatório, a demanda de energia da internet juntamente com as empresas vinculadas à nuvem foi de aproximadamente 623 bilhões de KWh em 2007; se a nuvem fosse um país, sua demanda de energia seria a quinta maior

[3] MOORE, G. E. Cramming more components onto integrated circuits. *Electronics*, 38 (8), 1965, p. 114-117.

[4] GREENPEACE. *How clean is your cloud*. Disponível em: <http://www.greenpeace.org/international/en/publications/Campaign-reports/Climate-Reports/How-Clean-is-Your-Cloud/>. Acesso em: 2 abr. 2012.

do mundo. Além disso, há uma estimativa de que a quantidade de informação digital cresça 50 vezes até 2020 e atinja um investimento aproximado de meio trilhão de dólares.

Esse crescimento exponencial na quantidade de dados coletados e armazenados não se restringe à internet. As empresas, de maneira geral, também ampliaram significativamente suas bases de dados não apenas em virtude da própria capacidade de armazenagem e disponibilidade de dados, mas também por causa de um forte aumento na qualidade e na quantidade de sensores capazes de gerar e monitorar dados. Ao mesmo tempo que a maioria das organizações despende bastante tempo e esforço na construção e manutenção de bases de dados, o que gerou inclusive especialidades como os DBAs (*database administrators*) e negócios como as empresas de indexação de bancos de dados, na maioria das vezes o conhecimento contido nas bases de dados corporativas é subvalorizado ou subutilizado. Algumas bases crescem tanto que nem os administradores conhecem as informações que podem ser extraídas ou a relevância que elas podem ter para o negócio. Cabe ressaltar que frequentemente os dados não podem ser analisados manualmente em virtude de fatores como grande quantidade de registros, elevado número de atributos, valores ausentes, presença de dados qualitativos e não quantitativos, entre outros.

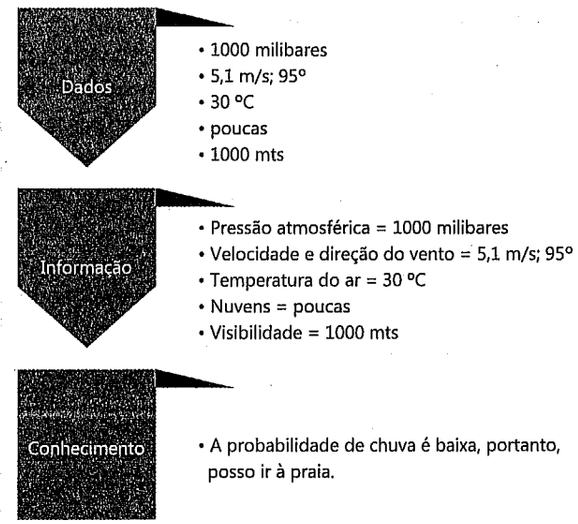
É exatamente nesse contexto de superabundância de dados que surgiu a *mineração de dados*, como um processo sistemático, interativo e iterativo, de preparação e extração de conhecimentos a partir de grandes bases de dados. Este livro tem como objetivo apresentar, de maneira didática, todo o processo de mineração de dados e as principais ferramentas de análise baseadas em técnicas de aprendizagem de máquina e alguns modelos de redes neurais. Os fundamentos matemáticos necessários a uma compreensão adequada do conteúdo do livro encontram-se resumidos no Apêndice I. Nas próximas seções deste capítulo, o leitor encontrará uma definição do que é mineração de dados, uma discussão sobre as diferentes nomenclaturas existentes na área de inteligência artificial como um todo e exemplos de aplicações práticas de mineração de dados.

1.2 O QUE É MINERAÇÃO DE DADOS?

O processo de *mineração* corresponde à extração de *minerais valiosos*, como ouro e pedras preciosas, a partir de uma *mina*. Uma característica importante desses materiais é que, embora não possam ser cultivados ou produzidos artificialmente, existem de maneira implícita e muitas vezes desconhecida em alguma fonte, podendo ser extraídos. Esse processo requer acesso à mina, o uso de ferramentas adequadas de mineração, a extração dos minérios propriamente dita e o seu posterior preparo para comercialização.

O termo *mineração de dados* (MD) foi cunhado como alusão ao processo de mineração descrito anteriormente, uma vez que se explora uma *base de dados* (mina) usando *algoritmos* (ferramentas) adequados para obter *conhecimento* (minerais preciosos). Os *dados* são símbolos ou signos não estruturados, sem significado, como valores em uma tabela, e a *informação* está contida nas descrições, agregando significado e utilidade aos dados, como o valor da temperatura do ar. Por fim, o *conhecimento* é algo que permite uma tomada de decisão para a agregação de valor, então, por exemplo, saber, que vai chover no fim de semana pode influenciar sua decisão de viajar ou não para a praia (Figura 1.2).

Figura 1.2 Exemplo da diferença entre dados, informação e conhecimento



A mineração de dados é parte integrante de um processo mais amplo, conhecido como *descoberta de conhecimento em bases de dados* (*knowledge discovery in databases*, ou *KDD*). Embora muitos usem mineração de dados como sinônimo de KDD, na primeira conferência internacional sobre KDD, realizada na cidade de Montreal, Canadá, em 1995, foi proposto que a terminologia descoberta de conhecimentos em bases de dados se referisse a todo o processo de extração de conhecimentos a partir de dados. Foi proposto também que a terminologia mineração de dados fosse empregada exclusivamente para a etapa de descoberta do processo de KDD,[5] que inclui a *seleção e integração das bases de dados*, a *limpeza da base*, a *seleção e transformação dos dados*, a *mineração* e a *avaliação dos dados*.

Este livro sintetiza o processo de KDD em quatro partes principais (Figura 1.3):

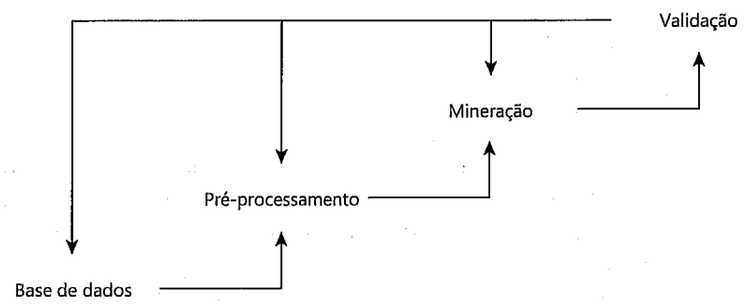
- **Base de dados:** coleção organizada de dados, ou seja, valores quantitativos ou qualitativos referentes a um conjunto de itens, que permite uma recuperação eficiente dos dados. Conceitualmente, os dados podem ser entendidos como o nível mais básico de abstração a partir do qual a informação e, depois, os conhecimentos podem ser extraídos (vide Figura 1.2);
- **Preparação ou pré-processamento de dados:** são etapas anteriores à mineração que visam preparar os dados para uma análise eficiente e eficaz. Essa etapa inclui a *limpeza* (remoção de ruídos e dados inconsistentes), a *integração* (combinação de

[5] ADRIAANS, P.; ZANTINGE, D. *Data mining*. Harlow: Addison-Wesley, 1996; HAN, J.; KAMBER, M.; PEI, J., *Data mining:* concepts and techniques. 3. ed. São Francisco: Morgan Kaufmann, 2011.

dados obtidos a partir de múltiplas fontes), a *seleção* ou *redução* (escolha dos dados relevantes à análise) e a *transformação* (transformação ou consolidação dos dados em formatos apropriados para a mineração);

- **Mineração de dados:** essa etapa do processo corresponde à aplicação de algoritmos capazes de extrair conhecimentos a partir dos dados pré-processados. Serão discutidas técnicas de *análise descritiva* (medidas de distribuição, tendência central e variância, e métodos de visualização), *agrupamento* (segmentação de bases de dados), *predição* (*classificação* e *estimação*), *associação* (determinação de atributos que coocorrem) e *detecção de anomalias*; e
- **Avaliação ou validação do conhecimento:** avaliação dos resultados da mineração objetivando identificar conhecimentos verdadeiramente úteis e não triviais.

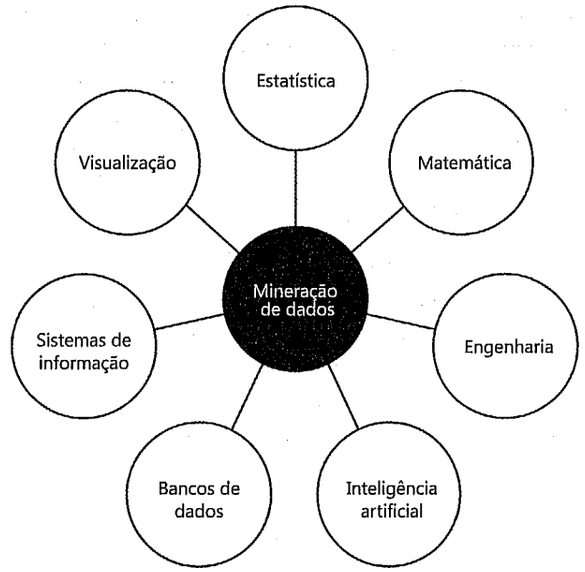
Figura 1.3 Processo de descoberta de conhecimento em bases de dados



Essas quatro etapas são correlacionadas e interdependentes de tal forma que a abordagem ideal para extrair informações relevantes em bancos de dados consiste em considerar as inter-relações entre cada uma dessas etapas e sua influência no resultado final. O processo de mineração de dados deverá permitir que conhecimentos interessantes e úteis sejam extraídos da base de dados e validados sob diferentes perspectivas. Esse conhecimento poderá ser usado para a tomada de decisão estratégica, como controle de processos, gestão da informação e conhecimento, processamento de consultas e muitas outras aplicações.

Cabe ressaltar que, sob uma perspectiva de armazém de dados (*data warehouse*), o processo de mineração de dados pode ser visto como um estágio avançado do processamento analítico *on-line* (*on-line analytical processing – OLAP*). Entretanto, a mineração de dados vai muito além do escopo restrito típico de um OLAP, baseado em métodos de resumo ou sumarização de dados, incorporando técnicas mais avançadas para a compreensão e a extração de conhecimentos dos dados.

Figura 1.4 Multidisciplinaridade da mineração de dados



A mineração de dados é uma disciplina interdisciplinar e multidisciplinar que envolve conhecimento de áreas como banco de dados, estatística, aprendizagem de máquina, computação de alto desempenho, reconhecimento de padrões, computação natural, visualização de dados, recuperação de informação, processamento de imagens e de sinais, análise espacial de dados, inteligência artificial, entre outras. A Figura 1.4 apresenta, de forma não exaustiva, algumas das principais áreas envolvidas na mineração de dados. O foco de apresentação deste livro é as técnicas de aprendizagem de máquina e alguns modelos de redes neurais (outras técnicas, como aquelas baseadas em computação natural, são encontradas em livros específicos).

1.2.1 Principais tarefas da mineração de dados

As funcionalidades da mineração de dados são usadas para especificar os tipos de informações a serem obtidas nas tarefas de mineração. Em geral, essas tarefas podem ser classificadas em duas categorias: (1) *descritivas*: caracterizam as propriedades gerais dos dados; e (2) *preditivas*: fazem inferência a partir dos dados objetivando predições. Em muitos casos, o usuário não tem ideia do tipo de conhecimento contido nos dados ou como usá-lo para gerar modelos preditivos, tornando importante a capacidade das ferramentas de mineração em encontrar diferentes tipos de conhecimento. As principais tarefas de mineração de dados são descritas sucintamente nesta seção e detalhadamente em capítulos dedicados ao longo do texto.

Análise descritiva de dados

Os algoritmos de aprendizagem de máquina são ferramentas poderosas para a descoberta de conhecimentos em bases de dados. Entretanto, uma etapa inicial do processo de mineração que não requer elevado nível de sofisticação é a *análise descritiva dos dados*, ou seja, o uso de ferramentas capazes de medir, explorar e descrever características intrínsecas aos dados. Especificamente, essas análises permitem investigar a *distribuição de frequência*, as *medidas de centro e variação*, e as *medidas de posição relativa e associação* dos dados. Além disso, técnicas elementares de *visualização* também são empregadas para um melhor entendimento da natureza e distribuição dos dados.

As análises descritivas permitem uma sumarização e compreensão dos objetos da base e seus atributos, como qual o salário médio dos professores universitários brasileiros ou qual a distribuição salarial desses professores. Usando essas medidas, é possível saber, por exemplo, qual a posição relativa de um salário quando comparada à distribuição de salários disponível, o que permite identificar, por sua vez, se um salário está abaixo ou acima da média. Essas informações podem ser representadas por meio de gráficos do tipo torta, gráficos em barra, histogramas ou outras ferramentas equivalentes, cada uma capaz de explicitar um conhecimento específico sobre os dados.

Predição: classificação e estimação

Predição é uma terminologia usada para se referir à construção e ao uso de um modelo para avaliar a classe de um objeto não rotulado ou para estimar o valor de um ou mais atributos de dado objeto. No primeiro caso, denominamos a tarefa de *classificação* e, no segundo, denominamos de *regressão* (em estatística) ou simplesmente *estimação*. Sob essa perspectiva, classificação e estimação constituem os dois principais tipos de problemas de predição, sendo que a classificação é usada para predizer *valores discretos*, ao passo que a estimação é usada para predizer *valores contínuos*.

Para exemplificar, considere o problema de atribuição de crédito, no qual um cliente se dirige a uma instituição financeira com o objetivo de conseguir um financiamento para trocar seu veículo. A primeira pergunta a ser respondida corresponde a uma tarefa de classificação: o crédito será oferecido ou não? Em seguida, há outra pergunta que pode ser relevante responder: qual o valor do crédito a ser oferecido? Essa última é uma estimação e ela faz sentido na medida em que o sistema de predição percebe que o cliente possui uma capacidade de pagamento superior ao que está sendo solicitado ou que o valor solicitado é muito alto, mas pode ser ajustado à sua capacidade financeira. Nesse caso, uma ferramenta capaz de estimar a capacidade de pagamento do cliente pode gerar maior lucro ou segurança para a empresa financiadora.

Como os rótulos das classes dos dados de treinamento são conhecidos *a priori* e usados para ajustar o modelo de predição, esse processo é denominado *treinamento supervisionado* (ou *aprendizagem supervisionada*). Exemplos de tarefas de classificação incluem identificação de *spams*, classificação de objetos, atribuição de crédito e detecção de fraudes. Exemplos de tarefas de estimação incluem predição de produtividade de grãos, estimativa de desempenho de atletas, estimativa de crédito, estimativa de valores futuros em bolsas de valores e previsão do clima.

Análise de grupos

Agrupamento (*clustering*) é o nome dado ao processo de separar (particionar ou segmentar) um conjunto de objetos em *grupos* (do inglês *clusters*) de objetos similares. Diferentemente da tarefa de classificação, o agrupamento de dados considera dados de entrada não rotulados, ou seja, o grupo (classe) ao qual cada dado de entrada (objeto) pertence não é conhecido *a priori*. O processo de agrupamento (ou *clusterização*) normalmente é utilizado para identificar tais grupos e, portanto, cada grupo formado pode ser visto como uma classe de objetos. Como os rótulos das classes dos dados de treinamento não são conhecidos *a priori*, esse processo é denominado *treinamento não supervisionado* (ou *aprendizagem não supervisionada*).

Em um processo de agrupamento, os objetos são agrupados com o objetivo de maximizar a distância interclasse e minimizar a distância intraclasse, ou, dito de outra forma, maximizar a similaridade intraclasse e minimizar a similaridade interclasse. Portanto, um *cluster* pode ser definido como uma coleção de objetos similares uns aos outros e dissimilares aos objetos pertencentes a outros *clusters*.

Para ilustrar uma tarefa de agrupamento, considere o problema de segmentar uma base de dados descrevendo frutas, na qual cada fruta está descrita por um conjunto de atributos, como forma, cor e textura. Suponha que haja maçãs e bananas nessa base de dados e que o algoritmo precisa segmentá-los sem ter conhecimento algum sobre a classe da fruta, recebendo apenas informações dos atributos. Como a forma, cor e textura das bananas são substancialmente diferentes da forma, cor e textura das maçãs, durante o agrupamento o algoritmo deverá, naturalmente, colocar bananas em um grupo e maçãs em outro.

Associação

Nas análises de grupos e preditivas, o objetivo em geral é encontrar relações (grupos, classes ou estimativas) entre os objetos da base. Entretanto, há diversas aplicações práticas nas quais o objetivo é encontrar relações entre os atributos (ou variáveis), e não entre os objetos. Para ilustrar esse caso, vamos considerar uma aplicação típica em marketing: a análise de carrinho de supermercado. Nesse tipo de análise, há um conjunto de transações (pedidos ou compras), e o objetivo é encontrar itens (produtos) que são comprados em conjunto; nesse sentido, as transações correspondem aos objetos da base e os itens, aos atributos.

Para ficar mais claro, pense no seguinte: os gerentes de marketing gostam muito de frases como "90% dos clientes que compram um *smartphone* assinam um plano de dados para seu aparelho"; nesse caso, a regra encontrada pela ferramenta de análise de dados e que está refletida nessa afirmação é aquela que associa *smartphone* ao plano de dados. Regras dessa natureza são chamadas de *regras de associação*.

A *análise por associação*, também conhecida por *mineração de regras de associação*, corresponde à descoberta de regras de associação que apresentam valores de atributos que ocorrem concomitantemente em uma base de dados. Esse tipo de análise costuma ser usado em ações de marketing e para o estudo de bases de dados transacionais. Há dois aspectos centrais na mineração de regras de associação: a proposição ou *construção* eficiente das

regras de associação e a quantificação da *significância* das regras propostas. Ou seja, um bom algoritmo de mineração de regras de associação precisa ser capaz de propor associações entre itens que sejam estatisticamente relevantes para o universo representado pela base de dados.

Mais formalmente, regras de associação possuem a forma $X \longrightarrow Y$:

$$A_1 \text{ e } A_2 \text{ e } \ldots \text{ e } A_m \longrightarrow B_1 \text{ e } B_2 \text{ e } \ldots \text{ e } B_n,$$

onde A_i, $i = 1, \ldots, m$, e B_j, $j = 1, \ldots, n$, são pares de valores de atributos.

As regras de associação $X \longrightarrow Y$ são interpretadas da seguinte forma: registros da base de dados que satisfazem à condição em X também satisfazem à condição em Y. No caso do exemplo apresentado no início desta seção, X = *smartphone* e Y = plano de dados. Além disso, outro aspecto importante mencionado no exemplo é que "*smartphone* $\longrightarrow$ plano de dados" em 90% dos casos, ou seja, há uma *confiança* de 90% de que um cliente que comprar um *smartphone* também assinará um plano de dados. Essa informação é estratégica para o negócio, pois pode induzir, por exemplo, promoções conjuntas de *smartphones* e planos de dados. A confiança é uma medida de significância da regra.

Considere agora outro exemplo de regra de associação que afirma que pessoas com idade igual ou superior a 16 anos e que possuem computador acessam redes sociais:

$$\text{idade } (X, \geq 16) \text{ e computador } (X, \text{SIM}) \longrightarrow \text{acessa } (X, \text{rede social})$$

onde X é uma variável que representa uma pessoa.

Suponhamos que essa regra ocorra com uma frequência igual a 20% da base e que em 80% das vezes que essa regra aparece ela seja verdadeira. Nesse caso, dizemos que a regra possui *cobertura*, ou *suporte*, igual a 20% e *confiança*, ou *acurácia*, igual a 80%. As medidas de cobertura e confiança são comuns para a análise de significância das regras geradas, e, nesse sentido, o objetivo do algoritmo de mineração de regras de associação é encontrar regras que satisfaçam critérios mínimos de significância. Em aplicações de marketing, por exemplo, esse tipo de metodologia pode ser usado para identificar quais itens são comprados em conjunto; já na detecção de fraudes, essa metodologia pode permitir a identificação de características ou comportamentos que ocorrem simultaneamente em uma fraude.

Detecção de anomalias

Uma base de dados pode conter objetos que não seguem o comportamento ou não possuem a característica comum dos dados ou de um modelo que os represente. Esses dados são conhecidos como *anomalias* ou *valores discrepantes* (*outliers*). A maioria das ferramentas de mineração descarta as anomalias – por exemplo, ruídos ou exceções –, entretanto, em algumas aplicações, como na detecção de fraudes, os eventos raros podem ser mais informativos do que aqueles que ocorrem regularmente.

As anomalias podem ser detectadas de diversas formas, incluindo métodos estatísticos que assumem uma distribuição ou modelo de probabilidade dos dados, ou medidas de distância por meio das quais objetos substancialmente distantes dos demais são considerados anomalias. Por exemplo, no caso de fraudes em cartões de crédito, valores muito

acima dos usuais para um dado cliente, assim como o tipo, o local e a frequência de uma dada compra, são indicativos de uma possível anomalia.

Uma característica marcante das anomalias é que elas compõem uma classe que ocorre com uma frequência bem inferior à(s) da(s) classe(s) normal(is). Isso faz com que os algoritmos de classificação e suas respectivas medidas de avaliação sejam fortemente impactados, forçando o uso de algoritmos e medidas de desempenho desenvolvidos especificamente para tratar tais problemas. Ao mesmo tempo, a vasta amplitude de problemas nessa área e sua relevância prática motivam a discussão em separado do tema.

1.2.2 Dicas para uma análise eficiente e eficaz

A mineração pode levar a uma capacidade preditiva e analítica poderosa dos dados. Mesmo quando aplicada de maneira correta, a capacidade de trabalhar com múltiplas variáveis e suas relações pode tornar os processos de mineração e interpretação dos resultados substancialmente complexos. Considerando essa complexidade, é preciso que o *analista de dados*, também conhecido como *cientista de dados*, esteja atento aos fundamentos conceituais necessários para o uso e o entendimento de cada técnica. A seguir, apresentamos uma lista de considerações (inevitavelmente incompleta) que podem servir como guia para uma mineração eficiente e eficaz:

- **Estabelecer a significância da mineração:** tanto a significância estatística quanto a prática da mineração devem ser consideradas. A significância estatística está relacionada à confiabilidade dos resultados obtidos, ou seja, se a base de dados foi preparada adequadamente para a análise, se os resultados apresentados são coerentes e se os algoritmos propostos tem o desempenho desejado. Por exemplo, uma amostragem ou normalização inadequada da base pode gerar resultados que não tenham nenhuma significância estatística e que, portanto, são inúteis. A significância prática, por sua vez, questiona sobre a aplicabilidade prática das análises realizadas, ou seja, se essas análises podem ser usadas em algum processo de tomada de decisão.
- **Reconhecer que as características da base de dados influenciam todos os resultados:** o processo de mineração opera, quase em sua totalidade, sobre uma base de dados pré-processada. Assim, é importante reconhecer que a quantidade de objetos na base, a dimensão (número de atributos) desses objetos, o tipo de atributos e seus domínios, a ausência de valores na base, as inter-relações entre os atributos e muitas outras características dos dados afetarão fortemente o resultado da análise, podendo, inclusive, invalidá-la.
- **Necessidade de conhecer os dados:** a discussão apresentada implica que análises preliminares dos dados – mais especificamente as técnicas de análise descritiva, como medidas de tendência central (por variável), análise de componentes principais e muitos outros métodos (estatísticos) simples – devem ser aplicados à base com o objetivo de entendê-la melhor antes de se iniciar a mineração propriamente dita.

- **Buscar pela parcimônia:** boa parte dos algoritmos de mineração resulta em uma espécie de modelo dos dados que poderá ser utilizado posteriormente para fazer alguma inferência ou predição. É possível que a escolha de diferentes amostras dos dados, ou mesmo diferentes execuções dos algoritmos, resultem em modelos com características distintas. Nesses casos, a escolha por um ou outro modelo deve considerar, entre outros aspectos, a parcimônia da solução, ou seja, a complexidade do modelo resultante. Muitas vezes, a complexidade de criação do modelo é um aspecto crucial na escolha de uma ferramenta dentro de um conjunto de possibilidades.
- **Verificar os erros:** todos os algoritmos de mineração podem ter seu desempenho avaliado. No caso dos algoritmos de agrupamento, há medidas que permitem avaliar a qualidade dos agrupamentos propostos; nas tarefas de predição, é possível avaliar o erro de predição; na mineração de regras de associação, avalia-se a significância das regras; e, para os algoritmos de detecção de anomalias, verifica-se o seu desempenho por meio de medidas específicas para esse tipo de problema. Em todos os casos, é preciso fazer um diagnóstico de desempenho do algoritmo, identificando os erros, o porquê de sua ocorrência, e empregar esse conhecimento para realimentar o processo de análise.
- **Validar seus resultados:** os resultados de uma análise precisam ser validados de diversas formas, por exemplo, comparando com o resultado de outras técnicas, analisando a capacidade de generalização dos métodos, combinando com outras técnicas e até utilizando um especialista de domínio capaz de validar se os resultados apresentados fazem sentido e se são de boa qualidade. Assim como no caso anterior, a validação é central para realimentar o processo de análise de dados.

1.3 AS DIFERENTES NOMENCLATURAS

A literatura está permeada por diferentes nomenclaturas para as muitas técnicas de solução de problemas e algoritmos computacionais que surgiram nas últimas décadas, e esse arsenal de métodos vem sendo desenvolvido e aplicado por diferentes pesquisadores, grupos de pesquisa, empresas, consultores e até pessoas comuns, utilizando os mais variados recursos teóricos, práticos, computacionais ou fontes de inspiração, desde a estatística até fenômenos só observados na natureza.

Essa quantidade de envolvidos e técnicas faz com que naturalmente surjam nomenclaturas distintas para contextos muitas vezes comuns, causando confusão e dificuldade de entendimento. Dentre as muitas nomenclaturas disponíveis na literatura técnico-científica merecem destaque as seguintes: *inteligência artificial, inteligência computacional, aprendizagem de máquina* e *computação natural*. Além dessas, uma nova terminologia, intimamente relacionada ao conteúdo deste livro, chamada de *big data*, vem sendo amplamente usada, sobretudo no mundo empresarial. As próximas seções fazem uma breve descrição do significado de cada uma dessas nomenclaturas, justificando, em alguns casos, o porquê de sua proposição.

1.3.1 Inteligência artificial clássica

J. McCarthy, um dos pioneiros da *inteligência artificial* (IA), define a área como a ciência e engenharia de máquinas inteligentes, especialmente programas inteligentes de computador. Ela está relacionada à tarefa de usar computadores para entender a inteligência humana, mas se restringindo necessariamente aos métodos inspirados na biologia.[6] Outra definição muito usada para a IA é aquela apresentada no livro de S. Russel e P. Norvig:[7] eles definem a IA como uma tentativa de entender e construir entidades inteligentes, e uma razão para estudá-la é aprender mais sobre nós mesmos. Nota-se que o foco dessas definições e a fonte básica de inspiração para o desenvolvimento da IA era a inteligência humana, nossa capacidade de percepção, resolução de problemas, comunicação, aprendizagem, adaptação e muitas outras.

As técnicas mais tradicionais de inteligência artificial, que surgiram na década de 1950 e prevaleceram até a década de 1980, ficaram conhecidas como *IA clássica*. Elas eram essencialmente *simbólicas*, ou seja, propunham que uma manipulação algorítmica de estruturas simbólicas – por exemplo, palavras – seria necessária e suficiente para o desenvolvimento de sistemas inteligentes. Essa tradição simbólica engloba também as abordagens baseadas em *lógica*, nas quais os símbolos são utilizados para representar objetos e relações entre objetos, e estruturas simbólicas são utilizadas para representar fatos conhecidos.

Uma característica marcante da IA clássica era a forma utilizada para construir o sistema inteligente. Existia uma visão procedural sugerindo que sistemas inteligentes poderiam ser projetados codificando-se conhecimentos especialistas em algoritmos específicos. Esses sistemas foram denominados genericamente *sistemas baseados em conhecimento* (*knowledge-based systems*) ou *sistemas especialistas* (*expert systems*). Um exemplo clássico de sistema especialista é para diagnóstico médico, em que a ideia central é que se faça uma representação simbólica do conhecimento do médico acerca de um estudo específico e, a partir de então, o diagnóstico é feito com base na relação das regras representadas nesse modelo (Figura 1.5).

Figura 1.5 Exemplo de sistema especialista para diagnóstico médico

> **Se** tosse **e** garganta inflamada **e** nariz escorrendo **e** febre, **então** gripe.
>
> **Se** cansaço **e** dor de cabeça **e** febre **e** feridas vermelhas na pele, **então** catapora.

Atualmente, a IA clássica envolve basicamente os sistemas especialistas, diversos métodos de busca – como busca em profundidade e busca em largura –, alguns sistemas

[6] McCARTHY, J. *What is artificial intelligence?*, Stanford University, 2007. Disponível em: <http://www.formal.stanford.edu/jmc/whatisai/whatisai.html>. Acesso em: 5 mar. 2012.

[7] RUSSEL, S. J.; NORVIG, P. *Artificial intelligence:* a modern approach. 3. ed. Upper Saddle Rive: Prentice Hall, 2009.

baseados em agentes e sistemas de raciocínio ou inferência baseados em lógica. Este livro não trata de nenhuma dessas técnicas em particular, concentrando-se naquelas de uma área denominada *aprendizagem de máquina*, a ser discutida mais adiante.[8]

1.3.2 Inteligência computacional

A proposta da inteligência artificial clássica era bastante ousada: projetar máquinas e organismos inteligentes capazes de realizar as mais diversas tarefas, desde fritar um ovo até dirigir veículos e se comunicar fluentemente com os humanos. Essa ambição foi relatada inclusive pela indústria cinematográfica em filmes como *Eu, Robô*; *2001: Uma Odisseia no Espaço*; *2010: O Ano em que Fizemos Contato*; *Blade Runner* e nas séries *Star Trek*; *O Exterminador do Futuro*; *Inteligência Artificia*; *Ela* e muitos outros. Entretanto, a dificuldade encontrada pela IA clássica em prover suas promessas (robôs inteligentes, veículos autoguiados etc.), gerou várias discordâncias entre ela e as abordagens que tinham essencialmente outras formas de operar, como as *redes neurais artificiais*, os *sistemas nebulosos* (*fuzzy systems*) e os *algoritmos evolutivos*. Um dos motivos principais dessa discordância era a disputa por financiamentos.

Houve então uma necessidade de dissociar essas áreas das técnicas que compunham a IA clássica e, para isso, criou-se uma nova linha de pesquisa denominada *inteligência computacional*. A primeira vez em que esses três grupos de técnicas foram apresentados em conjunto e se consolidaram como área de pesquisa foi no Congresso Mundial de Inteligência Computacional (World Congress on Computational Intelligence – WCCI), realizado em 1994 na cidade de Orlando, Estados Unidos, e que deu origem ao primeiro livro da área.[9] Desde então, o maior evento da área no mundo passou a ocorrer a cada quatro anos, tendo sua segunda edição realizada na cidade de Anchorage, Estados Unidos, em 1998, a terceira edição em Honolulu, Estados Unidos, em 2002, e a quarta edição em Vancouver, Canadá, em 2006, quando passou a ocorrer bienalmente. As edições seguintes deixaram os Estados Unidos e foram para a Ásia, a Europa e a Austrália.[10]

1.3.3 Aprendizagem de máquina

Em seu livro pioneiro, T. Mitchell[11] define *aprendizagem de máquina* (AM) como a área de pesquisa que visa desenvolver programas computacionais capazes de automaticamente melhorar seu desempenho por meio da experiência. A área de AM está baseada em conceitos

[8] O leitor interessado em aprofundar seus conhecimentos em IA pode consultar a seguinte bibliografia: RUSSEL, S. J.; NORVIG, P. *Artificial intelligence:* a modern approach. 3. ed. Upper Saddle Rive: Prentice Hall, 2009; LUGER, G. F. *Artificial intelligence:* structures and strategies for complex problem solving. 5. ed. UK: Addison Wesley, 2004; NILSSON, N. J. *Artificial intelligence:* a new synthesis. São Francisco: Morgan Kaufmann Publishers, 1998.

[9] ZURADA, J. M.; MARKS II, R. J.; ROBINSON, C. J. *Computational intelligence:* imitating life. Piscataway: IEEE Press, 1994.

[10] O leitor interessado em conhecimentos na área pode consultar a seguinte bibliografia: ZURADA, J. M.; MARKS, II. R. J.; ROBINSON, C. J. (1994); KONAR, A. *Computational intelligence:* principles, techniques and applications. Berlim: Springer, 2005; EBERHART, R. C.; SHI, Y. *Computational intelligence:* concepts to implementations. São Francisco: Morgan Kaufmann, 2007; ENGELBRECHT, A. P. *Computational intelligence:* an introduction. 2. ed. Nova York: Wiley, 2007; e RUTKOWSKI, L. *Computational intelligence:* methods and techniques. Heidelberg: Springer, 2010.

[11] MITCHELL, T. M. *Machine learning*. Nova York: McGraw-Hill, 1997.

e resultados de muitas outras áreas, como estatística, inteligência artificial, filosofia, teoria da informação, biologia, ciências cognitivas, complexidade computacional e teoria de controle. Seguindo uma linha similar, Alpaydin[12] define a aprendizagem de máquina como a programação de computadores para otimizar um critério de desempenho usando experiências passadas, chamadas de exemplos ou simplesmente dados de entrada. A ideia é que as técnicas envolvidas na AM sejam capazes, de alguma forma, de *aprender a resolver os problemas*.

Sistemas que sofrem aprendizagem são aqueles capazes de se *adaptar* ou *mudar seu comportamento* com base em exemplos, de forma que manipule informações. Duas virtudes importantes da aprendizagem baseada em adaptação são a possibilidade de resolver tarefas de processamento de informação e a capacidade de operar em ambientes dinâmicos. A maioria dos processos de aprendizagem é gradativa, ou seja, a aprendizagem não ocorre instantaneamente, requerendo um processo interativo e/ou iterativo de adaptação e interação com o ambiente.

Quando um sistema aprende alguma coisa, ele altera seu padrão comportamental ou alguma outra de suas características. Existem formas de aprendizagem que não são gradativas, como a memorização, e é importante salientar que a aprendizagem não requer consciência nem inteligência. Animais e insetos aprendem os caminhos que devem seguir para obter comida, reproduzir, construir suas casas e se proteger contra predadores. Aqueles malsucedidos nessas tarefas normalmente não sobrevivem em um ambiente com recursos limitados dentro do qual há uma luta pela sobrevivência.

A aprendizagem de máquina tem como foco extrair informação a partir de dados de maneira automática. Portanto, ela está intimamente relacionada à mineração de dados, à estatística, à inteligência artificial e à teoria da computação, além de outras áreas como computação natural, sistemas complexos adaptativos e computação flexível, como veremos a seguir. Os principais métodos investigados em aprendizagem de máquina são aqueles que trabalham com dados nominais, como as *árvores de decisão*, as *regras de associação* e *classificação*, tabelas de decisão e outros. Além desses, destacam-se os algoritmos baseados na *Teoria de Bayes*, alguns *métodos estatísticos* e *métodos de agrupamento de dados*.[13]

1.3.4 Paradigmas de aprendizagem

A capacidade de *aprender* associada às técnicas de aprendizagem de máquina é uma das mais importantes qualidades dessas estruturas. Trata-se da habilidade de *adaptar-se* ao ambiente de acordo com regras preexistentes, alterando seu desempenho ao longo do tempo. Assim, considera-se *aprendizado* o processo que adapta o comportamento e conduz a uma melhoria de desempenho de acordo com critérios preestabelecidos.

No contexto de mineração de dados, *aprendizagem* ou *treinamento* corresponde ao processo de ajuste e/ou construção do modelo usando um mecanismo de apresentação ou uso dos objetos da base de dados. Por exemplo, em uma árvore de decisão, o treinamento consiste em escolher atributos da base de dados que comporão cada nível de nós da

[12] ALPAYDIN, E. *Introduction to machine learning*. 2. ed. Cambridge: MIT Press, 2009.

[13] Muitas dessas técnicas serão vistas em detalhes neste livro e o leitor interessado pode buscar mais conteúdo sobre esses assuntos na seguinte bibliografia: MITCHELL (1997); ALPAYDIN (2009); MARSLAND, S. *Machine learning:* an algorithmic perspective. Boca Raton: CRC Press, 2009; ROGERS, S.; GIROLAMI, M. *A first course in machine learning*. Boca Raton: CRC Press, 2011.

árvore e construir os ramos de forma que otimize algum critério de qualidade; no algoritmo de agrupamento das *k*-médias, o treinamento consiste em apresentar os objetos da base de dados e ajustar a posição de um conjunto de vetores, chamados protótipos, que representam os grupos de objetos da base.

Há casos em que a aprendizagem só ocorre no momento do uso do sistema e, portanto, não é realizado um ajuste ou construção prévia do modelo. Normalmente, o algoritmo armazena toda a base de dados e a usa para inferir algo a respeito dos novos objetos dos quais se deseja obter alguma informação, como classe a que pertencem – esse tipo de aprendizagem é denominado *aprendizagem preguiçosa* (*lazy learning*). Um exemplo de algoritmo que opera dessa forma é o algoritmo dos *k* vizinhos mais próximos (*k*-NN, do inglês *k Nearest Neighbors*). O *k*-NN opera da seguinte maneira: dado um objeto cuja classe se deseja conhecer, esse objeto é comparado com todos os objetos da base de dados e sua classe é tomada como aquela dos *k* objetos mais próximos (similares) a ele.

Um procedimento bem definido para treinar uma técnica de aprendizagem de máquina é denominado *algoritmo de aprendizagem* ou *algoritmo de treinamento*, e a maneira pela qual o ambiente influencia a técnica em seu aprendizado define o *paradigma de aprendizagem*. Os dois paradigmas de aprendizagem mais comuns e que serão amplamente explorados neste livro são:

- **Aprendizado supervisionado:** é baseado em um conjunto de objetos para os quais as saídas desejadas são conhecidas, ou em algum outro tipo de informação que represente o comportamento que deve ser apresentado pelo sistema;
- **Aprendizado não supervisionado:** é baseado apenas nos objetos da base, cujos rótulos são desconhecidos. Basicamente, o algoritmo deve aprender a "categorizar" ou rotular os objetos.

1.3.5 Computação natural

Em meados dos anos 1960, novos sistemas começaram a ser desenvolvidos pela observação de outros fenômenos inteligentes naturais além da inteligência humana. Por exemplo, quem classificaria o mecanismo utilizado pelos cupins para a construção de seus ninhos como um comportamento inteligente? A partir desse mesmo princípio, vários outros exemplos podem ser inspirados na natureza, como evolução das espécies, construção de colmeias de abelhas, coleta de comida por formigas, entre outros.

Em um trabalho importante de formalização da área, Castro[14] propôs que a *computação natural* (CN) é uma terminologia introduzida para descrever três classes de métodos: 1) aqueles inspirados na natureza para o desenvolvimento de novas técnicas de solução de problemas; 2) aqueles baseados no uso de computadores para sintetizar fenômenos naturais; e 3) aqueles que fazem uso de materiais naturais (por exemplo, moléculas) para computar. De forma similar, Kari e Rozenberg[15] definem a computação natural como a linha de

[14] de CASTRO, L. N. Fundamentals of natural computing: an overview. *Physics of Life reviews*, 4, 2007, p. 1-36.
[15] KARI, L.; ROZENBERG, G. The many facets of natural computing, *Communications of the ACM*, n. 51, v. 10, 2008, p. 72-83.

pesquisa que investiga modelos e técnicas computacionais inspiradas na natureza e, dualmente, tenta compreender o mundo sob a perspectiva de processamento de informação.

Assim como todas as outras áreas discutidas até aqui, a computação natural é multidisciplinar e envolve conceitos de matemática, computação, estatística, biologia, química, engenharia e física. Ela se diferencia das demais por estar fundamentada numa relação próxima entre natureza e computação. Redes neurais artificiais, algoritmos evolutivos, *sistemas imunológicos artificiais*, *sistemas endócrinos artificias*, algoritmos baseados em *inteligência de enxame*, projetos de *vida artificial*, *geometria fractal*, *computação com moléculas* e *computação quântica*, todas fazem parte da CN, e cada uma dessas subáreas é, por si só, uma ampla área de pesquisa. A computação natural propõe um guarda-chuva conceitual comum para abranger todo esse amplo espectro de metodologias que unem natureza e computação.[16]

1.4 EXEMPLOS DE APLICAÇÃO

Há uma vasta literatura sobre aplicações de técnicas de mineração de dados em problemas nas mais variadas áreas. São típicas aplicações como análise e predição de crédito, detecção de fraudes, predição do mercado financeiro, relacionamento com clientes, predição de falência corporativa e muitas outras. Exemplos de segmentos de aplicação incluem setor financeiro; planejamento estratégico empresarial; planejamento do setor portuário; setores de energia (petróleo, gás, energia elétrica, biocombustíveis etc.); educação; logística; planejamento das cadeias de produção, distribuição e suprimentos; meio ambiente; e internet (portais, redes sociais, comércio eletrônico etc.). Aplicações típicas incluem identificação ou segmentação de clientes, parceiros, colaboradores; detecção de fraudes e anomalias em sistemas e processos; ações estratégicas de marketing, CRM e RH; jogos e atividades educacionais; gestão do conhecimento; análise de padrões de consumo; compreensão de bases de dados industriais, biológicas, empresariais e acadêmicas; predição de retorno sobre investimento, despesas, receitas, investimentos etc.; e mineração de dados da web. Esta seção lista com um pouco mais de detalhes alguns exemplos práticos de aplicação de mineração de dados.

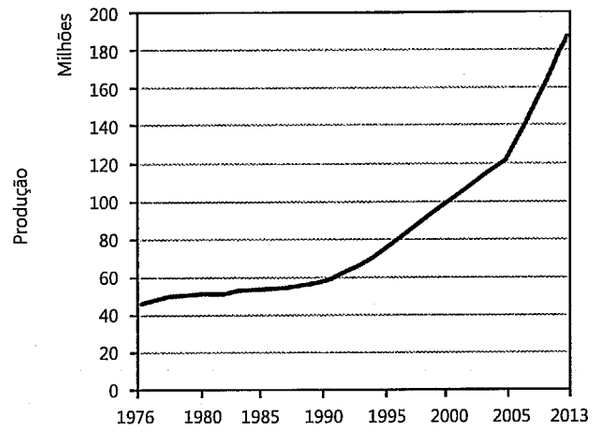
1.4.1 Predição de produtividade de grãos

Com relação a valor econômico, o Brasil é o quarto maior exportador de produtos agropecuários do mundo, ficando atrás apenas da União Europeia (composta por 25 países), seguida dos Estados Unidos da América e Canadá. Depois do Brasil estão a China, a Austrália, a Tailândia e a Argentina. Nossa produção de grãos vem crescendo vertiginosamente ao longo da história, passando de 46.943 mil toneladas em 1976-1977 para 188.658 mil

[16] Mais informações sobre a área podem ser encontradas nos textos propostos por de CASTRO, L. N. *Fundamentals of natural computing:* basic concepts, algorithms, and applications. Boca Raton: CRC Press, 2006; FLOREANO, D.; MATTIUSSI, C. *Bio-inspired artificial intelligence:* theories, methods, and technologies. Cambridge: The MIT Press, 2008; ZOMAYA, A. Y. *Handbook of nature-inspired and innovative computing*. Nova York: Springer, 2010; e ROZENBERG, G.; BACK, T.; KOK, J. N. *Handbook of natural computing*. Nova York: Springer, 2012.

toneladas em 2012-2013[17] (Figura 1.6). Apesar do crescente aumento da produção de grãos no país, o setor ainda sofre com altas dívidas, baixo custo do grão, como matéria-prima, quando comparada a produtos industrializados, infraestrutura deficiente e pouco uso de tecnologia.

Figura 1.6 Crescimento da produtividade de grãos no Brasil desde a década de 1970



Algoritmos de estimação podem ser utilizados para prever a produtividade de grãos em lavouras, o que é muito importante principalmente em um país que ainda possui boa parte de sua balança comercial equilibrada pela exportação de produtos agrícolas. Estimar, ou seja, prever o resultado de uma colheita pode ajudar na indicação de técnicas para a correção do solo, adequação dos processos de irrigação e melhoria do controle de pragas, tudo isso feito antes da colheita e, portanto, evitando possíveis prejuízos financeiros e até ambientais.

Uma maneira de empregar análise de dados para predição de colheita é utilizando amostras de folhas de plantas e amostras do solo da região para o treinamento dos algoritmos de predição. Para cada amostra de solo e folha da lavoura, podem-se obter as composições químicas com relação à concentração de alumínio, chumbo, potássio, magnésio, manganês e enxofre. Além dessas, também se pode considerar o pH do solo. O objetivo da predição é determinar a quantidade de calcário necessária para neutralizar o alumínio tóxico no solo, aumentar o cálcio, o magnésio e a base do solo. Com essas informações torna-se possível fazer a correção do solo antes da colheita.[18]

[17] Disponível em: <www.conab.gov.br>. Acesso em: 4 jan. 2016.

[18] de CASTRO, L. N.; VON ZUBEN, F. J.; MARTINS, W. Hybrid and constructive learning applied to a prediction problem in agriculture. *International Joint Conference on Neural Networks*, 3, 1998, p. 1932-1936.

1.4.2 Análise de sentimento em redes sociais

O poder da interação interpessoal em ambientes virtuais vem direcionando o mercado e promovendo a criação de novas empresas de internet. Consequentemente, inúmeros projetos de pesquisa e desenvolvimento surgiram com o objetivo de dar suporte a essas redes sociais tanto sob o ponto de vista tecnológico, quanto de modelo de negócios. Redes sociais como Flickr, Myspace, Facebook, LinkedIn, Twitter e muitas outras explodiram em popularidade nos últimos anos, também impulsionadas pelo aumento do poder aquisitivo da população mundial e pelas melhorias e redução de custos de toda a infraestrutura de comunicação.

Embora haja uma discussão se o Twitter é uma rede social ou uma mídia de informação,[19] ele é um serviço popular de microblog por meio do qual os usuários podem escrever mensagens curtas, chamadas tuítes, com até 140 caracteres, seguir outros usuários e ser seguidos. Essa possibilidade de estabelecer conexões entre pessoas é uma das características diferenciadoras das redes sociais.

Em meados de 2013 eram gerados, em média, cerca de 500 milhões de tuítes por dia,[20] carregados de opiniões sobre os mais diferentes assuntos, úteis para inteligência de marketing, psicólogos sociais, comércio eletrônico, monitoramento de reputação e muitos outros interessados na extração e mineração de opiniões, visões, humores e atitudes. A análise de dados do Twitter e de outras redes sociais pode, portanto, evidenciar por que determinados eventos repercutem na população. A aplicação de técnicas de mineração de dados possibilita extrair informações escondidas nos dados, como dispersão de doenças, posicionamento de candidatos a uma eleição, informações sobre catástrofes, monitoramento de marcas e muitas outras informações úteis e indispensáveis para a tomada de decisão estratégica.

Dentre as técnicas aplicáveis pode-se destacar a classificação de textos, que busca rotular um documento de acordo com suas características. Esse processo está inserido dentro do contexto da mineração de textos. A *análise de sentimento*, também conhecida como *mineração de opinião*, é um tipo de classificação de textos que objetiva rotulá-los de acordo com o sentimento ou a opinião neles contidos.[21,22] Classificar um texto de acordo com o sentimento que o usuário desejou passar, por exemplo, *positivo*, *negativo* ou *neutro*, permite o dimensionamento do retorno sobre determinado produto, serviço, marca, empresa etc. Para citar alguns poucos exemplos, consumidores podem usar a análise de sentimento para pesquisar sobre determinado produto ou serviço, empresas de marketing podem mensurar a opinião pública sobre uma campanha e empresas podem analisar críticas em uma nova versão de seu produto. É comum atualmente encontrarmos empresas especializadas no monitoramento e gerenciamento de redes sociais.

1.4.3 Detecção de fraudes em cartões de crédito

Atualmente, vários tipos de transações comerciais ocorrem quase em sua totalidade via cartões de crédito, como é o caso, por exemplo, de compras feitas pela internet e em lojas

[19] KWAK, H. et. al. What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on the World Wide Web* (WWW 2010), 2010, p. 591-600.

[20] Disponível em: <blog.twitter.com>. Acesso em: 4 jan. 2016.

[21] CAMBRIA, E. et. al. New avenues in opinion mining and sentiment analysis, *IEEE Intelligent Systems*, v. 28, n.2, 2013, p. 15-21.

[22] LIMA, A. C. E. S; de CASTRO, L. N.; CORCHADO, J. M. A polarity analysis framework for Twitter messages. *Applied Mathematics And Computation*, v. 270, 2015, p. 756-767.

de comércio eletrônico. O governo norte-americano estimou que, no final do século passado, aproximadamente 13 bilhões de dólares foram gastos em compras pela internet apenas usando cartões de crédito. O medo de transportar dinheiro no bolso e a crescente quantidade de estabelecimentos que aceitam cartões de crédito contribuem para uma utilização massiva de cartões por empresas e cidadãos comuns. Transações fraudulentas, conhecidas como *fraudes*, com cartões de crédito constituem um grande problema para a economia mundial, afetando desde as empresas que administram os cartões até os usuários legítimos dos cartões de crédito. O controle eficiente de transações comerciais feitas com cartões de crédito requer mecanismos de verificação e autenticação rápidos e eficazes que permitam um fácil uso pelos usuários legítimos, ao mesmo tempo em que garantam a detecção de tentativas de fraudes.

As fraudes em cartões de crédito podem ser divididas em duas grandes categorias: *fraudes comportamentais* e *fraudes de aplicação*.[23] As fraudes de aplicação ocorrem quando um indivíduo adquire um novo cartão de crédito utilizando informações pessoais falsas e, em seguida, gasta o máximo que pode em um curto intervalo de tempo. As fraudes comportamentais, por outro lado, são aquelas que ocorrem quando os detalhes (dados) de um usuário legítimo são obtidos e usados de forma fraudulenta; ou seja, transações ilegítimas são autorizadas sem ser detectadas pelas administradoras. As fraudes comportamentais podem ser resultado da interceptação de cartões de crédito enviados pelo correio, pela perda ou pelo roubo de um cartão, ou simplesmente pela aquisição e uso não autorizado de dados de um usuário legítimo.

No combate às fraudes, as ações da empresa administradora também podem ser agrupadas em duas grandes categorias: *prevenção* e *detecção*. A prevenção consiste em medidas que visam impedir a ocorrência de fraudes, como a necessidade de uso de senhas pessoais e sistemas de segurança para transações via web. Em contrapartida, a detecção de fraudes envolve a identificação rápida e eficiente de transações ilegítimas. É possível realizar a detecção de fraudes usando informações tanto sobre o padrão de comportamento normal de um usuário legítimo quanto usando dados sobre fraudes. Mesmo assim, a detecção de fraudes é um problema altamente complexo e desafiador em virtude de uma série de características:

- a quantidade de transações que são feitas diariamente é muito alta;
- o padrão de comportamento de um usuário legítimo pode mudar repentinamente (por exemplo, em viagens);
- a quantidade de transações legítimas é muito superior à quantidade de transações ilegítimas;
- os fraudadores adaptam constantemente seus comportamentos de acordo com a sofisticação dos sistemas de detecção; e
- diferentes transações envolvem diferentes quantias e, portanto, representam variáveis perdas potenciais.

[23] BOLTON, R. J.; HAND, D. J. Statistical fraud detection: a review. *Statistical Science*, v. 17, n. 3, 2002, p. 235-255; PHUA, C.; LEE, V.; GAYLER, R. A comprehensive survey of data mining-based fraud detection research, 2010. Disponível em: <http://arxiv.org/abs/1009.6119>. Acesso em: 13 set. 2015.

1.4.4 Combate a perdas não técnicas de energia elétrica

A existência de perdas em um sistema de energia elétrica é consequência natural do consumo de energia. As perdas podem ser categorizadas de acordo com o efeito, componente do sistema, ou causa e podem ser resumidas em:

- **Perdas técnicas:** correspondem àquelas perdas intrínsecas ao sistema elétrico, o que inclui as perdas nos equipamentos, na transformação e na distribuição da energia.
- **Perdas comerciais:** também chamadas de perdas não técnicas, são consequência, principalmente, de erros ou ausência de medição, medidores com defeito, consumidores clandestinos, desvio de consumo e furto de energia.

Um dos grandes problemas enfrentados pelas empresas distribuidoras de energia elétrica são as perdas comerciais provocadas intencionalmente por consumidores ou por falhas nos medidores. Diversos tipos de atividades têm sido aplicadas na redução dessas perdas, tais como campanhas publicitárias educativas, inspeções de consumidores, inspeções específicas em consumidores com perfil de consumo considerado suspeito, substituição de medidores eletromecânicos por medidores eletrônicos, programas de exteriorização da medição, operações de eliminação de ligações clandestinas, dentre outras.

Uma das formas de reduzir as perdas comerciais é realizar inspeções técnicas no local de consumo em busca de irregularidades, que vão desde a adulteração dos dispositivos de medição (fraude) até o furto ou desvio da energia propriamente dita. Entretanto, além da impossibilidade de inspecionar todos os consumidores, o custo associado à inspeção é alto, uma vez que esse processo demanda tempo, requer o deslocamento de uma equipe em campo e muitos dos consumidores inspecionados não são fraudadores. Com base nos dados de fiscalização obtidos a partir de medidas amostrais em campo, pode ser feita uma análise de dados para investigar inter-relações entre as amostras, segmentando os dados em grupos hierarquicamente vinculados, permitindo uma definição de pontos estratégicos de fiscalização.

Outra tarefa possível é a classificação automática dos cadastros disponíveis, a partir da qual se pode desenvolver um sistema de classificação que permita identificar de modo automático aqueles consumidores que provavelmente estejam causando perda de receita para a concessionária. Trata-se, portanto, de uma etapa na qual é feita a prospecção de possíveis perdas comerciais. Essa informação pode ser empregada no direcionamento das equipes de fiscalização e auditoria, impactando diretamente na redução das perdas não técnicas. Além dessas análises, dado o perfil de consumo dos usuários pode ser feito um levantamento das curvas típicas de hábito de consumo, permitindo uma identificação automática de novos clientes e de anomalias em clientes já existentes.

1.4.5 Segmentação de curvas de carga em sistemas de energia elétrica

Apesar do alto grau de desenvolvimento tecnológico da atualidade, só é possível armazenar energia elétrica em pequenas quantidades utilizando, para isso, baterias. No caso da energia elétrica consumida pelas indústrias, empresas e residências, a capacidade

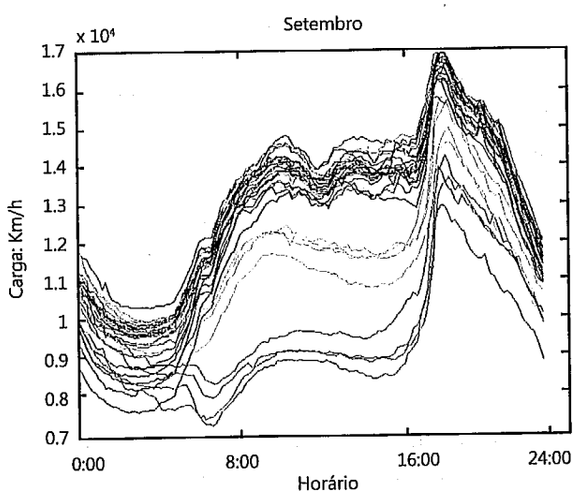
produtiva das usinas deve ser aproximadamente igual à quantidade de energia consumida. A pergunta que as usinas geradoras precisam responder, portanto, é qual será o consumo de energia elétrica a cada dia. Nesse contexto, é necessário realizar a predição da demanda de energia elétrica para que uma quantidade suficiente seja produzida.

A falta de planejamento e de investimentos no setor produtivo de energia elétrica pode causar apagões, cortes indesejáveis no fornecimento de energia, podendo até paralisar a produção industrial e deteriorar o desempenho de outros serviços. No Brasil três grandes apagões ocorreram nos anos 2001, 2002 e 2009 em decorrência da falta de planejamento ou outros problemas na geração ou distribuição da energia, o que levou o governo a estimular o racionamento voluntário, promovendo a economia e penalizando o desperdício de energia elétrica.

Com o objetivo de melhorar o planejamento da produção de energia elétrica, é possível usar técnicas de análise de dados para a previsão de carga (consumo) em curto, médio e longo prazos de um sistema elétrico de potência. No caso específico do curto prazo, para prever as cargas horárias de um dia, o padrão de carga horária e as cargas máxima e mínima devem ser determinados. Suponha que o objetivo inicial seja identificar dias da semana com padrões de cargas horárias similares e, posteriormente, realizar a previsão de demanda do setor. A previsão de demanda de carga é um meio de fornecer informações para uma tomada de decisão criteriosa que proporciona economia e segurança no fornecimento de energia elétrica. Para isso, uma companhia elétrica precisa resolver vários problemas técnicos e econômicos no planejamento e controle da operação do sistema de energia elétrica.

Para criar um modelo de segmentação de padrões de carga em sistemas de energia elétrica, pode-se utilizar uma base de dados referente ao consumo diário em determinados períodos do ano, como ilustrado na Figura 1.7. Ao observarmos os perfis dessas curvas de

Figura 1.7 Curvas de carga (consumo) de energia elétrica ao longo de um mês



carga, notamos a existência de padrões típicos de consumo. Após aplicar algoritmos de mineração de dados e analisar os resultados, é possível perceber, por exemplo, que os dias da semana entre terça e sexta-feira possuem padrão similar entre si, assim como os domingos, os sábados e as segundas-feiras; portanto, há quatro categorias distintas de perfis de carga. Classificar os perfis de carga anteriormente à previsão permite uma previsão de demanda mais precisa e exemplifica o fato de que as técnicas a serem discutidas aqui podem ser usadas em conjunto para se atingir um objetivo final. Nesse caso, um algoritmo de agrupamento é empregado antes de um algoritmo de predição.

1.4.6 Modelagem de processos siderúrgicos

Boa parte das siderúrgicas está equiparada tecnologicamente, sendo o uso eficiente do conhecimento um diferencial importante. O principal desafio é alcançar a excelência operacional pelo uso de tecnologias baseadas na experiência dos processos adquirida pelas pessoas e indústrias. As indústrias siderúrgicas investem esforços no desenvolvimento de tecnologias e dispositivos capazes de aumentar a produtividade das usinas, como a *sublança a oxigênio*, uma ferramenta importante para o controle do processo de *conversores* (fornos basculantes que têm a função de transformar a matéria-prima em aço líquido). A sublança é basicamente utilizada para medir o teor de carbono e a temperatura do aço durante o sopro de oxigênio, além de permitir retirar uma amostra que é enviada ao laboratório para análise detalhada da composição química do aço. A medição e a amostragem são realizadas antes do final do sopro de oxigênio e modelos matemáticos baseados nessa informação são utilizados para estimar a composição química que será obtida e, assim, redundar em ações corretivas do processo produtivo.

Algoritmos de mineração de dados podem ser usados para prever os principais elementos químicos (carbono, manganês, fósforo e enxofre) da análise de final de sopro sem utilizar os resultados da amostra da sublança. Essa solução permite uma redução do tempo de espera entre o recebimento do resultado da análise do laboratório e a execução do modelo de vazamento e pesagem das ferroligas. Dessa forma, a solução antecipa o final do tratamento nos conversores, possibilitando uma padronização, continuidade e uniformidade da operação, reduzindo o tempo de tratamento no conversor e aumentando a produtividade.