



SISTEMA GERENCIADOR DE BANCO DE DADOS

AULA 5



Profª Vívian Ariane Barausse de Moura

CONVERSA INICIAL

O objetivo desta aula é introduzir conceitos pertinentes ao *data mining*, ou seja, à mineração de dados. Para isso, aprenderemos conceitos básicos e avançaremos para a relação do processo da mineração de dados com a descoberta do conhecimento. Estudaremos as fases do processo de descoberta do conhecimento e os tipos de conhecimentos gerados de acordo com a bibliografia indicada.

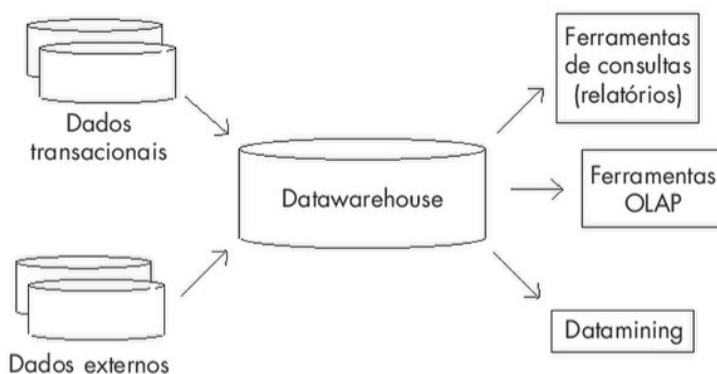
TEMA 1 – DATA MINING

De acordo com Ramakrishnan e Gehrke (2008, p. 737), *data mining*, ou “mineração de dados”,

consiste em encontrar tendências ou padrões interessantes em grandes conjuntos de dados para orientar decisões sobre atividades futuras. Há uma expectativa geral de que as ferramentas de mineração de dados deverão identificar esses padrões nos dados com entrada de usuário mínima. Os padrões identificados por essas ferramentas podem fornecer ao analista de dados ideias úteis e inesperadas que podem ser melhor investigadas subsequentemente, talvez se usando outras ferramentas de apoio à decisão.

Segundo Elsmari e Navathe (2011, p. 698), a mineração de dados está diretamente relacionada ao *data warehousing* (Figura 1), considerando que o “objetivo de um *data warehouse* é dar suporte à tomada de decisão com dados”. Assim, a mineração de dados pode ser usada junto com *data warehouse* para ajudar com certos tipos de decisões. A mineração de dados pode ser aplicada a banco de dados operacionais, com transações individuais; para que seja mais eficiente, o *data warehouse* deve ter uma coleção de dados agregada ou resumida.

Figura 1 – *Data warehouse* e outras tecnologias



Fonte: Caiçara, 2015, p. 197.

ElsMari e Navathe (2011) defendem que a mineração de dados auxilia na extração de novos padrões significativos que podem não ser necessariamente encontrados apenas ao consultar, processar dados ou metadados no *data warehouse*. Portanto, as aplicações e a mineração de dados devem ser fortemente consideradas desde o projeto de um *data warehouse*. Concomitantemente a isso, as ferramentas de mineração de dados devem ser projetadas para facilitar seu uso, juntamente com *data warehouse*, pois para bancos de dados muito grandes, que rodam *terabytes* ou até *petabytes* de dados, o uso bem-sucedido das aplicações de mineração de dados dependerá primeiro da construção de um *data warehouse*.

Segundo Elsmari e Navathe (2011, p. 700), os objetivos da mineração de dados estão diretamente relacionados com a descoberta do conhecimento. A mineração de dados costuma ser executada com objetivos finais ou aplicações de uma perspectiva geral. Esses objetivos se encontram nas seguintes classes: previsão, identificação, classificação e otimização, detalhadas no Quadro 1.

Quadro 1 – Objetivos da mineração de dados

Previsão	A mineração de dados pode mostrar como certos atributos dos dados se comportarão no futuro. Alguns exemplos de mineração de dados previsível incluem análise de transações de compra para prever o que os consumidores comprarão sobre certos descontos, quanto volume de vendas uma loja gerar ainda terminado o período e se exclusão de uma linha de produtos gerar mais lucros. Em tais aplicações, a lógica de negócios é usada junto com a mineração de dados. Em um contexto científico, certos padrões de ondas sísmicas podem prever um terremoto com alta probabilidade.
Identificação	Os padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade. Por exemplo, intrusos tentando quebrar um sistema podem ser identificados pelos programas executados, arquivos acessados e tempo de CPU por sessão. Em aplicações biológicas, a existência de um gênio pode ser identificada por certas sequências de símbolos nucleotídeos na sequência do DNA. Área conhecida como autenticação é uma forma de identificação. Ela confirma seu usuário realmente um usuário específico ou de uma classe autorizada, envolve uma comparação de parâmetros, imagens ou sinais contra um banco de dados.

(Continua)

Classificação	A mineração de dados pode particionar os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros. Por exemplo, os clientes em um supermercado podem ser categorizados em compradores que buscam desconto, compradores com pressa, compradores regulares leais, compradores ligados a marcas conhecidas e compradores eventuais. Essa classificação pode ser usada em diferentes análises de transações de compra de cliente como uma atividade pós-mineração. Às vezes, a classificação baseada em conhecimento de domínio comum é utilizada como uma entrada para decompor o problema de mineração e torná-lo mais simples. Por exemplo, alimentos saudáveis, alimentos de festa ou alimentos de lanche escolar são categorias distintas nos negócios do supermercado. Faz sentido analisar o relacionamento dentro e entre categorias como problemas separados. Essa categorização pode servir para codificar os dados corretamente Antes de submetê-los a mais mineração de dados.
Otimização	Um objetivo relevante na mineração de dados pode ser: utilizar o uso de recursos limitados, como o tempo, espaço, dinheiro ou materiais e maximizar variáveis de saída como vendas ou lucros sobre determinado conjunto de restrições. Como tal, esse objetivo da mineração de dados é semelhante à função objetiva, usada em problemas de pesquisa operacional, que lida com otimização sob restrições.

Fonte: Elaborado com base em Elsmari; Navathe, 2011, p. 700.

De acordo com Elsmari e Navathe (2011, p. 715) as tecnologias de mineração de dados podem ser aplicadas a uma grande variedade de contextos de tomada de decisão nos negócios. Em particular, algumas áreas de ganhos significativos, elencadas no quadro abaixo.

Quadro 2 – Aplicações de mineração de dados

Marketing	As aplicações incluem análise de comportamento do consumidor com base nos padrões de compra, determinação de estratégias de marketing (que incluem propaganda, local da loja e correio direcionado), segmentação de clientes, lojas ou produtos, e projetos de catálogos, além de layout de loja e campanhas publicitárias.
Finanças	As aplicações incluem análise de crédito de clientes, segmentação de contas a receber, análise de desempenho de investimentos financeiros, como ações, títulos e fundos de investimento, avaliação de opções de financiamento, e detecção de fraude.
Manufatura	As aplicações envolvem otimização de recursos como máquinas, mão de obra e materiais, e o projeto ideal de processos de manufatura, layout de galpões e projeto de produtos, como automóveis baseados em requisitos do cliente.
Saúde	Algumas aplicações são descobertas de padrões em imagens radiológicas e de análise de dados experimentais de <i>microarray</i> (chip de gene) para agrupar genes e relacionar sintomas ou doenças, para a análise de efeitos colaterais de drogas em eficácia de certos tratamentos, para a otimização de processos em um hospital e para relacionar dados de bem-estar do paciente com qualificações do médico.

Fonte: Elaborado com base em Elsmari; Navathe, 2011, p. 715.

ElsMari e Navathe (2011, p. 700) apontam que o termo *mineração de dados* é usado de uma forma popular e pode assumir um sentido muito amplo. Em algumas situações, inclui análise estatística e otimização restrita, bem como aprendizado de máquina, sendo que não há uma linha nítida separando a mineração de dados dessas disciplinas. Nesse sentido, os autores ressaltam que existem muitas particularidades a serem estudadas na aplicação de cada conceito.

TEMA 2 – MINERAÇÃO DE DADOS, CONHECIMENTO, ESTATÍSTICA E INTELIGÊNCIA COMPUTACIONAL

Ramakrishnan e Gehrke (2008, p. 738, grifo nosso) apontam que a mineração de dados “está relacionada à subárea da estatística chamada *análise de dados exploratória*, que tem objetivos semelhantes e conta com medidas estatísticas”. O autor explicita a sua relação com demais áreas de estudo, conforme abordado anteriormente. A mineração de dados também está intimamente relacionada às subáreas da inteligência artificial, como descoberta de conhecimento e aprendizado de máquina.

Uma das características relevantes que distingue a mineração de dados é que o volume tratado é muito grande. Ramakrishnan e Gehrke (2008) salientam que, embora as ideias de estudo relacionadas sejam aplicáveis aos problemas da mineração de dados, a capacidade de mudança de escala com relação ao tamanho dos dados é um novo critério importante. De acordo com os autores, “um algoritmo tem a capacidade de mudança de escala se o tempo corrente aumenta (linearmente) em proporção ao tamanho do conjunto de dados, mantendo constantes os recursos do sistema (por exemplo, quantidade de memória principal e velocidade de processamento da CPU) disponíveis”. (Ramakrishnan; Gehrke, 2008, p. 738). Nesse aspecto, algoritmos antigos devem ser adaptados ou novos algoritmos precisam ser desenvolvidos para garantir a capacidade de mudança de escala ao se descobrir padrões nos dados.

Encontrar tendências úteis em conjuntos de dados é uma definição bastante imprecisa da mineração de dados – de certo modo, pode-se considerar que todas as consultas de banco de dados fazem exatamente isso. Na verdade, temos ferramentas de análise e exploração com consultas SQL de um lado, consultas OLAP no meio e técnicas de mineração de dados do outro lado (Ramakrishnan; Gehrke, 2008).

As consultas SQL são construídas usando-se álgebra relacional – com algumas extensões, OLAP fornece idiomas de consulta de nível mais alto, com base no modelo de dados multidimensional, e a mineração de dados fornece operações de análise mais abstratas. Ramakrishnan e Gehrke (2008, p. 738) apontam que “podemos pensar nas diferentes tarefas de mineração de dados como ‘consultas’ complexas, especificadas em um alto nível, com alguns parâmetros definidos pelo usuário, e para as quais são implementados algoritmos especializados”.

Fazendo uma alusão com o mundo real, segundo Ramakrishnan e Gehrke (2008), a mineração de dados é muito mais do que simplesmente a aplicação de um desses algoritmos. Frequentemente, os dados contêm ruído ou gestão incompletos e, a não ser que isso seja entendido e corrigido, é provável que muitos padrões interessantes sejam perdidos e que a confiabilidade dos padrões detectados seja baixa. Além disso, o analista precisa decidir quais tipos de algoritmos de mineração são necessários, aplicá-los em um conjunto bem escolhido de amostras de dados e variáveis (isto é, tuplas e atributos), sintetizar resultados, aplicar outras ferramentas de apoio à decisão e mineração, e iterar o processo.

TEMA 3 – FASES DO PROCESSO DE DESCOBERTA DE CONHECIMENTO

A mineração de dados como parte do processo de descoberta do conhecimento é abordada, por Elsmari e Navathe (2011, p. 699), como “a descoberta de conhecimento nos bancos de dados”, que recebe a abreviação KDD (*Knowledge Discovery in Database*).

Segundo Ramakrishnan e Gehrke (2008, p. 739) o processo de descoberta de conhecimento e mineração de dados KDD pode ser separado em quatro etapas, conforme o Quadro 3.

Quadro 3 – Mineração de dados KDD

1. Seleção dos dados	O subconjunto objetivado dos dados e os atributos de interesse são identificados examinando-se o conjunto de dados bruto inteiro.
2. Limpeza dos dados	O ruído e exceções são removidos, valores de campo são transformados em unidades comuns e alguns campos são criados pela combinação de campos já existentes para facilitar a análise. Normalmente, os dados são colocados em um formato relacional e várias tabelas podem ser combinadas em uma etapa de desnormalização.
3. Mineração dos dados	Aplicamos algoritmos de mineração de dados para extrair padrões interessantes.
4. Avaliação	Os padrões são apresentados para os usuários finais em uma forma inteligível; por exemplo, por meio de visualização. O resultado de qualquer etapa no processo KDD pode nos levar de volta a uma etapa anterior para refazermos o processo com o novo conhecimento obtido. Neste capítulo, contudo, nos limitaremos a examinar algoritmos para algumas tarefas de mineração de dados específicas. Não discutiremos outros aspectos do processo KDD.

Fonte: Elaborado com base em Ramakrishnan; Gehrke, 2008, p. 739.

ElsMari e Navathe (2011) destacam que normalmente o processo KDD abrange mais do que a mineração de dados, pois que a descoberta de conhecimento compreende “seis fases: seleção de dados, limpeza de dados, enriquecimento, transformação ou codificação de dados, mineração de dados e o relatório em exibição da informação descoberta” (ElsMari; Navathe, 2011, p. 699).

Para exemplificar esse processo, os autores utilizam um exemplo: um banco de dados de transação, mantido por um revendedor de bens de consumo e especializados. Suponha que os dados do cliente incluam “o nome de cliente, CEP, número de telefone, data de compra, código do item, preço, quantidade e valor total. Uma grande quantidade de conhecimento novo pode ser descoberta pelo processamento KDD nesse banco de dados de cliente” (ElsMari; Navathe, 2011, p. 699), conforme exemplificado no Quadro 4.

Quadro 4 – Processamento KDD

1. Seleção de dados	Durante a seleção de dados, dados sobre itens específicos ou categorias de itens, ou de lojas em uma região ou área específica do país, podem ser selecionados.
2. Limpeza de dados	O processo de limpeza de dados, então, pode corrigir códigos postais inválidos ou eliminar registros com prefixos de telefone incorretos.
3. Enriquecimento	O enriquecimento normalmente melhora os dados com fontes de informação adicionais. Por exemplo, dados os nomes de cliente e números de telefone, a loja pode adquirir outros dados sobre idade, renda e avaliação de crédito e anexá-los a cada registro.
4. Transformação de dados	A transformação de dados e a codificação podem ser feitas para reduzir a quantidade de dados. Por exemplo, os códigos de item podem ser agrupados em relação a categorias de produtos, por exemplo: em áudio, vídeo, suprimentos, aparelhos eletrônicos, câmera, acessórios, e assim por diante. Os códigos postais podem ser agregados em regiões geográficas, às rendas podem ser divididas em faixas, e assim por diante.
5. Limpeza	A limpeza funciona como um precursor para criação do data warehouse, se a mineração de dados for baseada em um data warehouse existente para essa cadeia de varejo, podemos esperar que a limpeza já tenha sido aplicada. É somente depois do pré-processamento que as técnicas de mineração de dados são usadas para extrair diferentes regras e padrões.
6. Relatórios e exibição da informação descoberta	Os resultados da mineração de dados podem ser formados em diversos formatos, como listagem saídas gráficas tabelas de resumo, relatórios descritos ou visualizações.

Fonte: Elaborado com base em Elsmari; Navathe, 2011, p. 699.

A aplicação da mineração de dados pode gerar novos tipos de informação nova, apresentados no quadro abaixo.

Quadro 5 – Nova informação gerada

Regras de associação	Por exemplo: sempre que um cliente compra equipamento de vídeo ele ou ela também compra outro aparelho eletrônico.
Padrões sequenciais	Por exemplo: suponha que um cliente compre uma câmera e dentro de três meses ele ou ela compra e suprimentos fotográficos, depois, dentro de seis meses ele ou ela provavelmente comprar um item de acessório. Isso define um padrão sequencial de transações. Um cliente que compra mais que o dobro em períodos fracos provavelmente poderá comprar pelo menos uma vez durante o período de Natal.
Árvores de classificação	Por exemplo: os clientes podem ser classificados por frequência de visitas, tipos de financiamento utilizado, valor da compra ou afinidade para tipos de itens; algumas estatísticas reveladoras podem ser geradas para essas classes.

Fonte: Elaborado com base em Elsmari; Navathe, 2011, p. 699.

Conforme destacam Elsmari e Navathe (2011, p. 699) “existem muitas possibilidades para descobrir novos conhecimentos sobre padrões de compra, relacionando fatores como idade, grupo de renda, local de residência, o que e quanto os clientes compram”. Essas informações podem então ser utilizadas para planejar locais adicionais de loja, com base em demografia, e também realizar promoções, combinar itens em propagandas ou planejar estratégias de marketing sazonal.

Segundo os autores, esse exemplo de loja de varejo demonstra que a mineração de dados precisa ser precedida por uma preparação significativa nos dados, antes de gerar informações úteis que possam influenciar diretamente as decisões de negócios (Elsmari; Navathe, 2011).

TEMA 4 – TIPOS DE CONHECIMENTO DESCOBERTOS DURANTE A MINERAÇÃO DE DADOS

O termo *conhecimento* pode assumir diferentes aplicações. Em seus estudos, Elsmari e Navathe (2011, p. 700) afirmam que esse termo é interpretado de forma livre como algo que envolve algum grau de inteligência. Para que chegue nesse patamar, é necessário passar pela transformação de dados brutos, da informação para o conhecimento. Os autores classificam o conhecimento como dedutivo e indutivo.

Nesse espectro, “o conhecimento dedutivo deduz novas informações com base na aplicação de regras lógicas previamente especificadas de dedução sobre o dado indicado” (Elsmari; Navathe, 2011, p. 700).

Já o conhecimento indutivo descobre novas regras e padrões com base nos dados fornecidos, e é nesse aspecto que a mineração de dados enfoca. O conhecimento assim pode ser representado de várias maneiras: em um sentido desestruturado, pode ser representado por regras ou pela lógica proposicional; na forma estruturada, pode ser representado por árvores de decisão, redes semânticas, redes neurais e hierarquias de classes ou frames. Elsmari e Navathe (2011) defendem também que o conhecimento descoberto durante o processo de mineração de dados pode ser descrito de acordo com o quadro abaixo.

Quadro 6 – Conhecimento descoberto na mineração de dados

Regras de associação	Essas regras correlacionam a presença de um <i>itemset</i> (conjunto de itens) com outra faixa de valores para um conjunto de variáveis diverso. Exemplos: (1) quando uma compradora adquire uma bolsa, ela provavelmente compra sapatos; (2) Uma imagem de raio X contendo características a e b provavelmente também exibe a característica c .
Hierarquia de classificação	O objetivo é trabalhar partindo de um conjunto existente de eventos ou transações para criar uma hierarquia de classes. Exemplos: (1) Uma população pode ser dividida em cinco faixas de possibilidade de crédito com base em um histórico de transações de créditos anteriores; (2) Um modelo pode ser desenvolvido para os fatores que determinam o desejo de obter a localização de loja em uma escala de 1 a 10; (3) Companhias de investimentos podem ser classificadas com base nos dados de desempenho usando características como crescimento, receita e estabilidade.
Padrões sequências	Uma sequência de ações ou eventos é buscada. Exemplo: se um paciente passou por uma cirurgia de ponte de safena para artérias bloqueadas e um aneurisma e, depois, desenvolveu ureia sanguínea alta dentro de um ano da cirurgia, ele provavelmente sofrerá insuficiência renal nos próximos 18 meses. A detecção de padrões sequenciais é equivalente à detecção de associações entre eventos com certos relacionamentos temporais.
Padrões dentro da série temporal	As similaridades podem ser detectadas dentro de posições de uma série temporal de dados, que é uma sequência de dados tomados em intervalos regulares, como vendas diárias ou preços de ações de fechamento diário. Exemplos: (1) As ações de uma companhia de energia, ABC Power, e uma companhia financeira, XYZ Securities, mostraram o mesmo padrão durante 2009 em matéria de preços de fechamento de ações; (2) Dois produtos mostraram o mesmo padrão de vendas no verão, mas um padrão diferente no inverno; (3) Um padrão no vento magnético solar pode ser usado para prever mudanças nas condições atmosféricas da Terra.
Agrupamento	Determinada população de eventos ou itens pode ser particionada (segmentada) em conjuntos de elementos “semelhantes”. Exemplos: (1) Uma população inteira de dados de transação sobre uma doença pode ser dividida em grupos com base na similaridade dos efeitos colaterais produzidos; (2) A população adulta nos Estados Unidos pode ser categorizada em cinco grupos, desde mais prováveis de comprar até menos prováveis de comprar um novo produto; (3) Os acessos web feitos por uma coleção de usuários contra um conjunto de documentos (digamos, em uma biblioteca digital) podem ser analisados em relação às palavras-chave dos documentos para revelar grupos ou categorias de usuários.

Fonte: Elaborado com base em Elsmari; Navathe, 2011, p. 701.

TEMA 5 – FASES DO PROCESSO DE DESCOBERTA

Segundo Elsmari e Navathe (2011), uma das principais tecnologias em mineração de dados envolve a descoberta de regras de associação. Segundo Ramakrishnan e Gehrke (2008), muitos algoritmos foram propostos para descobrir várias regras que descrevem os dados sucintamente; nesse sentido, vamos apresentar alguns algoritmos relacionados. O banco de dados é

considerado uma coleção de transações, cada uma envolvendo um *itemset* (conjunto de itens).

Ramakrishnan e Gehrke (2008) tratam da relação apresentada na Figura 2 para ilustrar as regras de associação. A explicação dos itens da tabela segue em Ramakrishnan e Gehrke (2008, p. 704):

Examinando o conjunto de transações em Compras, podemos identificar regras da forma: Os registros aparecem ordenados em grupos, por transação. Todas as tuplas de um grupo têm o mesmo *idtrans* e, juntas, elas descrevem uma transação de cliente, a qual envolve compras de um ou mais *itens*. Uma transação ocorre em determinada data e o nome de cada item adquirido é registrado, junto com a quantidade comprada. Observe que há redundância em Compras: ela pode ser decomposta armazenando-se triplas *idtrans-idcli-data* em uma tabela separada e eliminando-se *idcli* e *data* de Compras; pode ser assim que os dados estejam realmente armazenados.

Figura 2 – Relação de compras

<i>idtrans</i>	<i>idcli</i>	<i>data</i>	<i>item</i>	<i>qtd</i>
111	201	5/1/99	caneta	2
111	201	5/1/99	tinta	1
111	201	5/1/99	leite	3
111	201	5/1/99	suco	6
112	105	6/3/99	caneta	1
112	105	6/3/99	tinta	1
112	105	6/3/99	leite	1
113	106	5/10/99	caneta	1
113	106	5/10/99	leite	1
114	201	6/1/99	caneta	2
114	201	6/1/99	tinta	2
114	201	6/1/99	suco	4
114	201	6/1/99	água	1

Fonte: Elaborado com base em Ramakrishnan; Gehrke, 2008, p. 740.

Usaremos a relação de compras da Figura 2 para ilustrar as regras de associação.

Examinando o conjunto de transações em compras, podem ser identificadas regras a partir da forma representada na Figura 3.

Figura 3 – Regra de associação

$$\{caneta\} \Rightarrow \{tinta\}$$

Fonte: Elaborado com base em Ramakrishnan; Gehrke, 2008, p. 744.

O autor indica que essa regra deve ser lida como se segue: “Se uma caneta é comprada em uma transação, é provável que tinta também seja comprada nessa transação” (Ramakrishnan; Gehrke, 2008, p. 744). Essa é uma afirmação que descreve as transações no banco de dados; a extrapolação para futuras transações deve ser feita com cuidado.

De maneira genérica, uma regra de associação tem a forma LHS→RHS, em que LHS (*left hand side*) e RHS (*right hand side*) são conjuntos de itens. A interpretação dessa regra é que, se cada item em LHS é comprado em uma transação, então, é provável que os itens em RHS também serão comprados. Segundo Ramakrishnan e Gehrke (2008), existem duas medidas importantes para uma regra de associação, apresentadas no quadro abaixo.

Quadro 7 – Medidas de regra de associação

Suporte	Confiança
O suporte de um conjunto de itens é a porcentagem de transações que contém todos esses itens. O suporte de uma regra LHS→RHS é o suporte do conjunto de itens LHS∪RHS. Por exemplo, considere a regra {caneta} → {tinta}. O suporte dessa regra é o suporte do conjunto de itens {caneta, tinta}, que é de 75%.	A confiança é com relação à implicação mostrada na regra. Considere as transações que contém todos os itens em LHS. A confiança de uma regra LHS→RHS é a porcentagem de tais transações que também contém todos os itens em RHS. Mais precisamente, seja sup (LHS) a porcentagem das transações que contém LHS e seja sup (LHS ∪ RHS) a porcentagem das transações que contém LHS e RHS. Então, a confiança da regra LHS → RHS é $\text{sup}(\text{LHS} \cup \text{RHS}) / \text{sup}(\text{LHS})$. A confiança de uma regra é a indicação de sua força. Como exemplo, considere novamente a regra {caneta} → {tinta}. A confiança dessa regra é de 75%; 75% das transações que contém o conjunto de itens {caneta} também contém o conjunto de itens {tinta}.

Fonte: Elaborado com base em Ramakrishnan; Gehrke, 2008, p. 721.

Ramakrishnan e Gehrke (2008) e Elsmari e Navathe (2011) detalham que um usuário pode solicitar todas as regras de associação que têm um suporte mínimo e confiança mínima especificados. Vários algoritmos foram desenvolvidos para encontrar tais regras eficientemente. De acordo com Ramakrishnan e Gehrke (2008), esses algoritmos procedem em duas etapas. Na primeira etapa, todos os conjuntos de itens frequentes com o suporte mínimo

especificado pelo usuário são computados. Na segunda, são geradas regras usando conjuntos de itens frequentes como entrada.

Alguns algoritmos citados são: algoritmo *a priori*, algoritmo de amostragem, algoritmo de árvore padrão frequente (FP) e de crescimento (FP) e algoritmo de partição (ElsMari; Navathe, 2011).

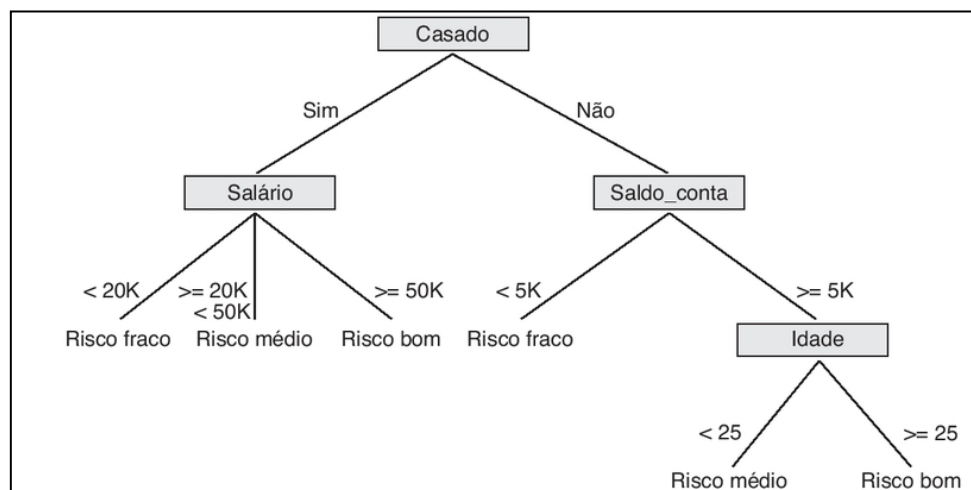
5.1 Classificação

De acordo com Elsmari e Navathe (2011), classificação é o processo de aprender um modelo que descreve diferentes classes de dados. As classes são pré-definidas utilizando o exemplo da aplicação bancária, quando os clientes que solicitam cartão de crédito podem ser classificados com risco fraco, risco médio ou risco bom. Os autores definem que esse tipo de atividade também é chamada de *aprendizado supervisionado*.

Quando o modelo é criado, ele pode ser usado para classificar novos dados; o primeiro passo é o do aprendizado do modelo, quando são realizados treinamentos de dados que já foram classificados. Cada registro nos dados de treinamento contém um atributo, chamado *rótulo de classe*, que indica a que classe o registro pertence. O modelo que é produzido costuma ser efetivado na forma de árvore de decisão ou conjunto de regras (ElsMari; Navathe, 2011).

Algumas das questões importantes com relação ao modelo e ao algoritmo que produz o modelo incluem a capacidade de o modelo prever a classe correta de dados novos, o custo computacional associado ao algoritmo e a sua escalabilidade. Elsmari e Navathe (2011) utilizam o modelo de uma árvore de decisão, que é a representação gráfica da descrição de cada classe – ou, em outras palavras, uma representação das regras de classificação conforme representado na Figura 3.

Figura 3 – Árvore de decisão da amostra para aplicações de cartão de crédito



Fonte: Elaborado com base em Elsmari; Navathe, 2011, p. 710.

A partir da Figura 3, Elsmari e Navathe (2011, p. 710) fazem a seguinte descrição: “se um cliente for casado e seu salário for maior ou igual a 50K, então ele tem um risco bom para um cartão de crédito bancário”. Essa é uma das regras que implicam no fato de a classe *risco bom* atravessar a árvore de decisão, saindo da raiz, para cada nó e folha, e formando outras regras para essa classe e as duas outras classes. Utilizando esse exemplo, inicialmente todas as amostras de treinamentos estão na raiz da árvore e são particionadas de maneira recursiva com base nos atributos selecionados. O atributo usado em um nó para particionar as amostras é aquele com o melhor critério de divisão – por exemplo, aquele que maximiza a medida de ganho da informação.

5.2 Padrões sequenciais

Segundo Elsmari e Navathe (2011, p. 713), a descoberta de padrões sequenciais é baseada “no conceito de uma sequência de conjunto de itens”. Consideramos que transações como a utilizada na Figura 2 (Relação de compras) são ordenadas no momento da compra. Essa ordenação gera uma sequência de *itemsets*. Por exemplo, {caneta, tinta}, {caneta, suco, água}, {leite, suco, água} podem ser uma sequência de *itemset* com base em três compras por um mesmo cliente. O suporte para uma sequência **S** de *itemsets* é a porcentagem do conjunto indicado **U** de sequências, em que **S** é uma subsequência. Nesse exemplo, {caneta, tinta}, {caneta, suco, água}, e {leite, suco, água} são consideradas *subsequências*. O problema de identificar padrões sequenciais, então, é encontrar todas as subsequências para os conjuntos de

sequências indicados que possuem o suporte mínimo definido pelo usuário. A sequência $S_1, S_2, S_3 \dots$ é um **indicador** do fato de que um cliente que compra os *itemsets* S_1 e depois S_2 provavelmente vai comprar o itemset S_3 , e assim por diante; essa previsão é baseada na frequência (suporte) dessa sequência no passado. Diversos algoritmos formaram foram investigados para detecção de sequência. (ElsMari; Navathe, 2011).

5.3 Padrões dentro da série temporal

ElsMari e Navathe (2011, p. 714) definem *séries temporais* como “sequências de eventos, cada evento pode ser um certo tipo fixo de uma transação”. Utilizam como exemplo o preço de fechamento de uma ação ou de um fundo, que é um evento que vai ocorrer a cada dia da semana para cada ação e fundo; a sequência desses valores por ação ou fundo constitui uma série temporal.

Para os autores, em uma série temporal pode-se procurar uma série de padrões ao analisar sequências e subsequentes. Poderíamos, por exemplo, achar o período durante o qual o preço da ação subiu ou se manteve constante por N dias, ou ainda achar o período mais longo sobre o qual o preço da ação teve uma flutuação de não mais do que 1% em relação ao preço de fechamento anterior. Os autores ainda citam que poderíamos procurar o trimestre durante o qual o preço da ação teve o maior ganho ou perda percentual (ElsMari; Navathe, 2011).

Para Elsmari e Navathe (2011), a série temporal pode ser comparada estabelecendo-se medidas de similaridade para identificar empresas cujas ações se comportam de modo semelhante. A “análise e a mineração de séries temporais” é uma funcionalidade estendida do gerenciamento de dados temporais.

5.4 Agrupamento

Para Ramakrishnan e Gehrke (2008, p. 754) “o objetivo do agrupamento é particionar um conjunto de registros em grupos tais que os registros dentro de um grupo sejam similares entre si e os registros pertencentes a dois grupos diferentes sejam diferentes”. Os autores indicam que cada grupo é chamado de *agrupamento* e cada registro pertence a exatamente um agrupamento.

Os autores destacam que a similaridade entre os registros é medida computacionalmente por uma função de distância. Uma função de distância recebe dois registros de entrada e retorna um valor que é uma medida de similaridade entre eles. Diferentes aplicações têm diferentes noções de similaridade e nenhuma medida funciona para todos os domínios. Normalmente, a saída de um algoritmo de agrupamento consiste em uma representação resumida de cada agrupamento. O tipo de representação resumida depende fortemente do tipo e do formato dos agrupamentos que o algoritmo calcula (Ramakrishnan; Gehrke 2008).

ElsMari e Navathe (2011, p. 712) defendem que o agrupamento “lida com o particionamento de dados com base no uso de uma mostra de treinamento pré-classificada”. Em geral, é útil particionar os dados sem ter uma amostra de treinamento – isso também é conhecido como aprendizado não supervisionado. Destacamos também que “o objetivo do agrupamento é colocar registros em grupos, de modo que os registros em um grupo sejam semelhantes uns aos outros e diferentes dos registros em outros grupos, os grupos costumam ser disjuntos” (ElsMari; Navathe, 2011, p. 712).

Cita-se, como exemplo, que no comércio pode ser importante determinar grupos de clientes que têm padrões de compras semelhantes, ou ainda, na medicina, pode-se determinar grupos de pacientes que mostram reações semelhantes aos medicamentos receitados. Para Elsmari e Navathe (2011), uma faceta importante do agrupamento é a função de similaridade em uso. Quando os dados são numéricos, normalmente é utilizada uma função de similaridade baseada na distância; os autores citam como exemplo a distância euclidiana, que pode ser usada para medir a similaridade.

FINALIZANDO

Nesta aula, estudamos sobre a mineração de dados, que se refere à descoberta de novas informações em termos de padrões ou regras com base em grandes quantidades de dados. A tecnologia de banco de dados é utilizada para descobrir conhecimento e padrões adicionais dos dados. Para realizar a mineração de dados, existem diversas técnicas, regras e algoritmos que podem ser empregados dentro das suas especificidades. As ferramentas de mineração de dados estão em uma evolução contínua e muitas delas são desenvolvidas com base em inteligência artificial, estatística e otimização.

REFERÊNCIAS

CAIÇARA, C. J. **Sistemas integrados de gestão: ERP** – uma abordagem gerencial. 2. ed. Curitiba: InterSaberes, 2015.

ELMASRI, R.; NAVATHE, S. B. **Sistema de banco de dados**. 6. ed. São Paulo: Pearson Education do Brasil, 2011.

RAMAKRISHNAN, R.; GEHRKE, J. **Sistemas de gerenciamento de bancos de dados**. 3. ed. Porto Alegre: McGraw Hill, 2008.