Faculdade de Tecnologia de São Paulo
Curso de Tecnologia em Processamento de Dados

Técnica de Árvore de Decisão em Mineração de Dados

São Paulo – SP
Dezembro de 2011

Faculdade de Tecnologia de São Paulo
Curso de Tecnologia em Processamento de Dados

Técnica de Árvore de Decisão em Mineração de Dados

Eric Ossamu Hosokawa

Monografia desenvolvida ao curso de Tecnologia em Processamento de Dados, da Faculdade de Tecnologia de São Paulo, como exigência parcial para obtenção do Título de Tecnólogo em Processamento de Dados. Sob a orientação da professora Sandra Harumi Tanaka.

São Paulo – SP
Dezembro de 2011

Resumo

Mineração de Dados é uma área de pesquisa multidisciplinar, incluindo tecnologia de bancos de dados, inteligência artificial, aprendizado de máquina, redes neurais, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados.

Uma árvore de decisão é uma forma gráfica de representar as decisões e suas possíveis conseqüências. Portanto, uma árvore de decisão nada mais é do que um mecanismo de ajuda na tomada de decisões. Este artigo mostra como se dá o seu funcionamento na mineração de dados.

Abstract

Data Mining is a multidisciplinary research area, including database tecnology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge based systems, information retrieval, high-performance computing and data visualization.

A decision tree is a graphic way to represent the decisions and their possible consequences. Therefore, a decision tree is nothing but a help mecanism to make decisions. This article shows how does this work on data mining.

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Sumário

1. Introdução

Mineração de Dados é um ramo da computação que teve início nos anos 80, quando os profissionais das empresas e organizações começaram a se preocupar com os grandes volumes de dados informáticos estocados e inutilizados dentro da empresa. Nesta época, Data Mining consistia essencialmente em extrair informação de gigantescas bases de dados da maneira mais automatizada possível. Atualmente, Data Mining consiste, sobretudo na análise dos dados após a extração, buscando-se, por exemplo, levantar as necessidades reais e hipotéticas de cada cliente para realizar campanhas de marketing. Assim, uma empresa de cartões de crédito, por exemplo, tem uma mina de ouro de informações: ela sabe os hábitos de compra de cada um dos seus seis milhões de clientes. O que costuma consumir, qual o seu padrão de gastos, grau de endividamento, etc. Para a empresa essas informações são extremamente úteis no estabelecimento do limite de crédito para cada cliente, além disso, contém dados comportamentais de compra de altíssimo valor. Os seguintes pontos são algumas das razões por que o Data Mining vem se tornando necessário para uma boa gestão empresarial: (a) os volumes de dados são muito importantes para um tratamento, utilizando somente técnicas clássicas de análise, (b) o usuário final não é necessariamente um estatístico, (c) a intensificação do tráfego de dados (navegação na Internet, catálogos online, etc) aumenta a possibilidade de acesso aos dados.

Este trabalho tem como objetivo fornecer um apanhado geral das principais tarefas e a técnicas de mineração de dados conhecida como Árvores de Decisão. Discutiremos algumas técnicas de otimização e implementação de algoritmos de mineração de dados referentes a tarefa de regras classificação/predição. Além disso, discutiremos aspectos teóricos subjacentes que possibilitarão o desenvolvimento de algoritmos de mineração para novas tarefas.

Em teoria, uma árvore de decisão é um gráfico em forma de árvore, contendo as decisões a serem tomadas e suas possíveis conseqüências (riscos, custo, prejuízos), usado para criar um plano para se alcançar um objetivo. Uma árvore de decisão é um modelo preditivo; Isto é, um mapeamento de observações sobre um item para conclusões sobre o seu valor-alvo. Cada nó interno corresponde uma variável; um arco para um nó-filho representa um possível valor daquela variável. Uma folha representa o valor previsto da variável-alvo, dadas as variáveis representadas no caminho até ela desde a raiz.

1.1 O Processo de Descoberta de Conhecimento em Bancos de Dados (KDD)

Considere-se uma hierarquia de complexidade: se algum significado especial é atribuído a um dado, ele se transforma em uma informação (ou fato). De acordo com Sade (1996), se uma norma (ou regra) é elaborada, a interpretação do confronto entre o fato e a regra constitui um conhecimento.

O processo KDD é constituído de várias etapas, como ilustrado na Figura 1.1, que são executadas de forma interativa e iterativa. De acordo com Brachman & Anand (1996), as etapas são interativas porque envolvem a cooperação da pessoa responsável pela análise de dados, cujo conhecimento sobre o domínio orientará a execução do processo. Por sua vez, a iteração deve-se ao fato de que, com freqüência, esse processo não é executado de forma seqüencial, mas envolve repetidas seleções de parâmetros e conjunto de dados, aplicações das técnicas de Data Mining e posterior análise dos resultados obtidos, a fim de refinar os conhecimentos extraídos.

Dentre as várias etapas do processo KDD, a principal, que forma o núcleo do processo e que, muitas vezes, confunde-se com ele, chama-se Data Mining.

Descoberta de conhecimento em bancos de dados é o processo não trivial de identificar em dados padrões que sejam válidos, novos (previamente desconhecidos), potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão. Examinando estes termos individualmente:

- Dados: conjunto de fatos *F*, como instâncias de um banco de dados. Por exemplo, uma coleção de *n* cadastros de pessoas físicas contendo: idade, profissão, renda etc.

- Padrão: expressão *E* em uma linguagem *L* descrevendo fatos em um subconjunto *FE* de *F*. *E* é dito um padrão se é mais simples do que a enumeração de todos os fatos em *FE*. Por exemplo, o padrão: “Se renda < \$*r* então a pessoa não recebe financiamento” seria aplicável para uma escolha apropriada de *r*.

- Processo: geralmente em KDD, processo é uma seqüência de vários passos que envolve preparação de dados, pesquisa de padrões, avaliação de conhecimento, refinamento envolvendo iteração e modificação.

- Validade: os padrões descobertos devem ser válidos em novos dados com algum grau de certeza. Uma medida de certeza é uma função *C* mapeando expressões em *L* para um espaço de medidas *MC* . Por exemplo, se um limite de padrão de crédito é ampliado, então a medida de certeza diminuiria, uma vez que mais financiamentos seriam concedidos a um grupo até então restrito a esta operação.

- Novo: em geral, assume-se que "novidade" pode ser medida por uma função *N(E,F)*, que pode ser uma função booleana ou uma medida que expresse grau de "novidade" ou "surpresa". Exemplo de um fato que não é novidade: sejam E = "usa tênis" e *F* = "alunos de colégio" então *N(E,F)* = 0 ou *N(E,F) = false.* Por outro lado: sejam *E* = "bom pagador" e *F* = "trabalhador da construção civil" então *N(E,F)* = 0,85 ou *N(E,F) = true*.

- Potencialmente útil: padrões devem potencialmente levar a alguma atitude prática, conforme medido por alguma função de utilidade. Por exemplo, regras obtidas no processo podem ser aplicadas para aumentar o retorno financeiro de uma instituição.

- Compreensível: um dos objetivos de KDD é tornar padrões compreensíveis para humanos, visando promover uma melhor compreensão dos próprios dados. Embora seja um tanto subjetivo medir compreensibilidade, um dos fatores freqüentes é a medida de simplicidade. O fator de compreensão dos dados está relacionado à intuitividade da representação destes, bem como da granularidade alta o suficiente para que estes sejam compreendidos. Por exemplo: o log de um servidor Web não é uma representação compreensível; já fatos estatísticos extraídos deste log, tais como totais de acesso ou classificação dos acessos realizados, fornecem informação num formato mais intuitivo e de granularidade humanamente compreensível.

1.2 O que é Mineração de Dados

Afinal, o que é Mineração de Dados? Falando simplesmente, trata-se de extrair ou minerar conhecimento de grandes volumes de dados. Muitas pessoas consideram o termo Mineração de Dados como sinônimo de *Knowledge Discovery in Databases* (KDD) ou Descoberta de Conhecimento em Banco de Dados, ou seja, Data Mining, ou Mineração de Dados, pode ser entendido como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomada de decisões. É uma metodologia aplicada em diversas áreas que usam o conhecimento, como empresas, indústrias e instituições de pesquisa.

Segundo Fayyad (1996), o processo de KDD é interativo, iterativo, cognitivo e exploratório, envolvendo vários passos com muitas decisões sendo feitas pelo analista (que é um especialista do domínio dos dados, ou um especialista de análise dos dados), na verdade, KDD é um processo mais amplo consistindo das seguintes etapas:

1. Definição do tipo de conhecimento a descobrir, o que pressupõe uma compreensão do domínio da aplicação, bem como do tipo de decisão que tal conhecimento pode contribuir para melhorar.

2. Criação de um conjunto de dados alvo (*Selection*): selecionar um conjunto de dados, ou focar num subconjunto, onde a descoberta deve ser realizada.

3. Limpeza de dados e pré-processamento (*Preprocessing*): operações básicas tais como: remoção de ruídos quando necessário, coleta da informação necessária para modelar ou estimar ruído, escolha de estratégias para manipular campos de dados ausentes, e formatação de dados de forma a adequá-los à ferramenta de mineração.

4. Redução de dados e projeção (*Transformation*): localização de características úteis para representar os dados dependendo do objetivo da tarefa, visando a redução do número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados, bem como o enriquecimento semântico das informações.

5. Mineração de dados (*Data Mining*): selecionar os métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação ou conjunto de representações; busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão.

6. Interpretação dos padrões minerados (*Interpretation/Evaluation*), com um possível retorno aos passos 1-6 para posterior iteração.

7. Implantação do conhecimento descoberto (*Knowledge*): incorporar este conhecimento ao desempenho do sistema, ou documentá-lo e reportá-lo às partes interessadas.

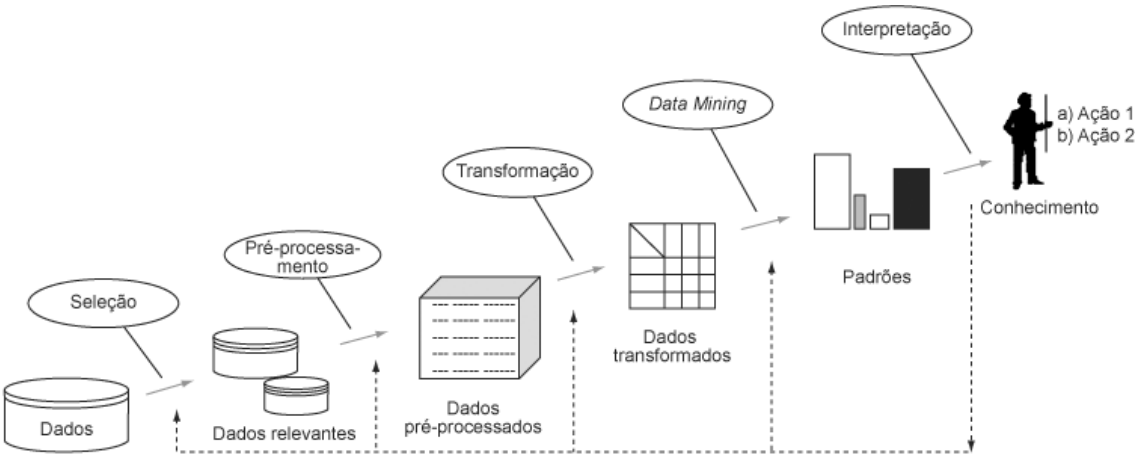

Figura 1.1: As etapas do processo de KDD (Fayyad, 1996).

Esta tarefa está focada, sobretudo na técnica freqüentemente utilizada na etapa Mineração do processo de KDD. Supomos que os dados já foram devidamente selecionados e transformados, integrados num armazém de dados (*Data Warehouse*) e deles foram eliminados ruídos que possam afetar o processo de descoberta de conhecimento. A fase de visualização do conhecimento descoberto também não é tratada neste trabalho.

Sendo assim, Data Mining define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para descrever características do passado como para predizer tendências para o futuro.

Para encontrar respostas ou extrair conhecimento interessante, existem diversas tarefas de Data Mining disponíveis na literatura. Mas, para que a descoberta de conhecimentos seja relevante, é importante estabelecer metas bem definidas. Essas metas são alcançadas por

meio das seguintes tarefas de Data Mining: Classificação, Modelos de Relacionamento entre Variáveis, Análise de Agrupamento, Sumarização, Modelo de Dependência, Regras de Associação e Análise de Séries Temporais, conforme citação e definição feita por Fayyad et al. (1996).

É importante ressaltar que a maioria dessas tarefas é baseada em técnicas das áreas de aprendizado de máquina, reconhecimento de padrões e estatística. Essas técnicas vão desde as tradicionais da estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos.

As etapas de mineração de dados utilizam técnicas e algoritmos de diferentes áreas do conhecimento, principalmente inteligência artificial (especialmente aprendizagem de máquina), banco de dados (recursos para manipular grandes bases de dados) e estatística na avaliação e validação de resultados).

Uma questão que freqüentemente surge é a seguinte: porque não utilizar tão somente os conhecidos procedimentos estatísticos para obter informações relevantes a partir de um conjunto de dados?

Conforme mencionado, procedimentos estatísticos são utilizados nas etapas de KDD e mais especificamente na mineração de dados. Entretanto, o volume, a complexidade e as peculiaridades dos eventos e dos dados por eles originados, impõem severas limitações a metodologias puramente estatísticas, dentre elas:

- Dados nem sempre possuem independência estatística entre eles, ou seja, muitos domínios possuem inter-relação entre seus objetos e respectivos atributos, comprometendo a aplicação de métodos estatísticos;

- A análise estatística demanda um grau de conhecimento e domínio desta área que apenas estatísticos e profissionais de áreas correlatas possuem, restringindo assim a atuação da grande maioria dos potenciais usuários de procedimentos analíticos;

- Métodos estatísticos manipulam muito bem dados numéricos, mas não manipulam bem valores simbólicos, incompletos ou inconclusivos;

- Estes métodos são computacionalmente caros quando se trata de grandes bases de dados.

Desta forma, percebe-se claramente que a mineração de dados possui grande relevância, contribuição e abrangência no que diz respeito a aplicações. Visando uma melhor compreensão das tarefas, será apresentado a seguir uma breve descrição dos principais métodos de mineração de dados, utilizando aprendizagem de máquina.

A exemplificação de cada tópico toma por base recursos do Weka, uma ferramenta de KDD que contempla uma série de algoritmos de preparação de dados, de aprendizagem de máquina (mineração) e de validação de resultados.

1.3 Tarefas de Mineração de Dados

É importante distinguir o que é uma tarefa e o que é uma técnica de mineração. A tarefa consiste na especificação do que estamos querendo buscar nos dados, que tipo de regularidades ou categoria de padrões temos interesse em encontrar, ou que tipo de padrões poderiam nos surpreender (por exemplo, um gasto exagerado de um cliente de cartão de crédito, fora dos padrões usuais de seus gastos). A técnica de mineração consiste na especificação de métodos que nos garantam como descobrir os padrões que nos interessam.

Dentre as principais técnicas utilizadas em mineração de dados, temos técnicas estatísticas, técnicas de aprendizado de máquina e técnicas baseadas em crescimento-poda validação. A seguir, descrevemos de forma sucinta as principais tarefas de mineração.

Classificação

Associa ou classifica um item a uma ou várias classes categóricas pré-definidas. Uma técnica estatística apropriada para classificação é a análise discriminante. Os objetivos dessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de várias populações, além da classificação das observações em uma ou mais classes predeterminadas.

A idéia é derivar uma regra que possa ser usada para classificar, de forma otimizada, uma nova observação a uma classe já rotulada. Segundo Mattar (1998), a análise discriminante permite que dois ou mais grupos possam ser comparados, com o objetivo de determinar se diferem uns dos outros e, também, a natureza da diferença, de forma que, com base em um conjunto de variáveis independentes, seja possível classificar indivíduos ou objetos em duas ou mais categorias mutuamente exclusivas.

Modelos de Relacionamento entre Variáveis

Modelos de Relacionamento entre Variáveis: associa um item a uma ou mais variáveis de predição de valores reais, consideradas variáveis independentes ou exploratórias. Técnicas estatísticas como regressão linear simples, múltipla e modelos lineares por transformação são utilizadas para verificar o relacionamento funcional que, eventualmente, possa existir entre duas variáveis quantitativas, ou seja, constatar se há uma relação funcional entre X e Y.

Observa-se, conforme Gujarati (2000), que o método dos mínimos quadrados ordinários, atribuído a Carl Friedrich Gauss, tem propriedades estatísticas relevantes e

apropriadas, que tornaram tal procedimento um dos mais poderosos e populares métodos de análise de regressão.

Análise de Agrupamento (Cluster)

Associa um item a uma ou várias classes categóricas (ou clusters), em que as classes são determinadas pelos dados, diversamente da classificação em que as classes são pré-definidas. Os clusters são definidos por meio do agrupamento de dados baseados em medidas de similaridade ou modelos probabilísticos.

A análise de cluster (ou agrupamento) é uma técnica que visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles. Nesse tipo de análise, segundo Pereira (1999), o procedimento inicia com o cálculo das distâncias entre os objetos estudados dentro do espaço multiplano constituído por eixos de todas as medidas realizadas (variáveis), sendo, a seguir, os objetos agrupados conforme a proximidade entre eles. Na seqüência, efetuam-se os agrupamentos por proximidade geométrica, o que permite o reconhecimento dos passos de agrupamento para a correta identificação de grupos dentro do universo dos objetos estudados.

Sumarização

Determina uma descrição compacta para um dado subconjunto. As medidas de posição e variabilidade são exemplos simples de sumarização. Funções mais sofisticadas envolvem técnicas de visualização e a determinação de relações funcionais entre variáveis. As funções de sumarização são freqüentemente usadas na análise exploratória de dados com geração automatizada de relatórios, sendo responsáveis pela descrição compacta de um conjunto de dados. A sumarização é utilizada, principalmente, no pré-processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas – como mínimo, máximo, média, moda, mediana e desvio padrão amostral –, no caso de variáveis quantitativas, e, no caso de variáveis categóricas, por meio da distribuição de freqüência dos valores.

Técnicas de sumarização mais sofisticadas são chamadas de visualização, que são de extrema importância e imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados. Exemplos de técnicas de visualização de dados incluem diagramas baseados em proporções, diagramas de dispersão, histogramas e box plots, entre outros. Autores como Levine et al. (2000) e Martins (2001), entre outros, abordam com grande detalhamento esses procedimentos metodológicos.

Modelo de Dependência

Descreve dependências significativas entre variáveis. Modelos de dependência existem em dois níveis: estruturado e quantitativo. O nível estruturado especifica, geralmente em forma de gráfico, quais variáveis são localmente dependentes. O nível quantitativo especifica o grau de dependência, usando alguma escala numérica.

Segundo Padovani (2000), análises de dependência são aquelas que têm por objetivo o estudo da dependência de uma ou mais variáveis em relação a outras, sendo procedimentos metodológicos para tanto a análise discriminante, a de medidas repetidas, a de correlação canônica, a de regressão multivariada e a de variância multivariada.

Regras de Associação

Determinam relações entre campos de um banco de dados. A idéia é a derivação de correlações multivariadas que permitam subsidiar as tomadas de decisão. A busca de associação entre variáveis é, freqüentemente, um dos propósitos das pesquisas empíricas. A possível existência de relação entre variáveis orienta análises, conclusões e evidenciação de achados da investigação. Uma regra de associação é definida como se X então Y, ou $X => Y$, onde X e Y são conjuntos de itens e $X \cap Y = 0$. Diz-se que X é o antecedente da regra, enquanto Y é o seu conseqüente. Medidas estatísticas como correlação e testes de hipóteses apropriados revelam a freqüência de uma regra no universo dos dados minerados.

Vários métodos para medir associação são discutidos por Mattar (1998), de natureza paramétrica e não paramétrica, considerando a escala de mensuração das variáveis.

Análise de Séries Temporais

Determina características seqüenciais, como dados com dependência no tempo. Seu objetivo é modelar o estado do processo extraindo e registrando desvios e tendências no tempo. Correlações entre dois instantes de tempo, ou seja, as observações de interesse, são obtidas em instantes sucessivos de tempo – por exemplo, a cada hora, durante 24 horas – ou são registradas por algum equipamento de forma contínua, como um traçado eletrocardiográfico. As séries são compostas por quatro padrões: tendência, variações cíclicas, variações sazonais e variações irregulares.

Há vários modelos estatísticos que podem ser aplicados a essas situações, desde os de regressão linear (simples e múltiplos), os lineares por transformação e regressões assintóticas, além de modelos com defasagem, como os auto-regressivos (AR) e outros deles derivados.

Uma interessante noção introdutória ao estudo de séries temporais é desenvolvida por Morettin & Toloi (1987).

Diante da descrição sumária de metodologias estatísticas aplicáveis ao procedimento de Mineração de Dados, registra-se que, embora Hand (1998) afirme que o termo Data Mining possa trazer uma conotação simplista para os estatísticos, Fayyad et al. (1996a) mostraram a relevância da estatística para o processo de extração de conhecimentos, ao afirmar que essa ciência provê uma linguagem e uma estrutura para quantificar a incerteza resultante quando se tenta deduzir padrões de uma amostra a partir de uma população.

De acordo com Hand (1998), a estatística preocupa-se com a análise primária dos dados, no sentido de que eles são coletados por uma razão particular ou por um conjunto de questões particulares a priori. Data Mining, por outro lado, preocupa-se também com a análise secundária dos dados, em um sentido mais amplo e mais indutivo do que uma abordagem hipotético-dedutiva, freqüentemente considerada como o paradigma para o progresso da ciência moderna. Assim, Data Mining pode ser visto como o descendente direto da estatística, já que são técnicas metodológicas complementares.

Técnica	Descrição	Tarefas	Exemplos
Árvore de Decisão	Baseada em estágios de decisão (nós) e na separação de classes e subconjuntos, organiza os dados de forma hierárquica.	-Classificação -Predição	CART, CHAID, C5.0, ID-3.
Redes Neurais	Modelos inspirados na fisiologia do cérebro, nos quais o conhecimento é fruto do mapa de conexões neuronais e dos pesos dessas conexões.	-Classificação -Agrupamento -Predição	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB.
Raciocínio Baseado em Casos	Baseado no método do vizinho mais próximo combina e compara atributos para estabelecer hierarquia de semelhança.	-Classificação -Agrupamento	BIRCH , CLARANS CLIQUE.
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, em que a cada nova geração, soluções melhores têm mais chance de ter "descendente".	-Classificação -Agrupamento	Algoritmo Genético Simples, Genitor, GA-Nuggets, GAPVMINER.
Conjuntos Fuzzy	Oferece uma grande vantagem para classificar dados com um alto nível de abstração.	-Classificação -Agrupamento	K-means. FCMdd
Regras de Indução	Processo para obter uma hipótese a partir de dados e fatos já existentes.	-Classificação -Predição	CART, CHAID
Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados.	-Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM.

Quadro 1.1: Técnicas e Tarefas utilizadas para a mineração de dados (GOLDSCHIMIDT, 2005).

1.4 Como Avaliar os Padrões Interessantes?

É muito importante que os resultados e modelos possam ser avaliados e comparados. Alguns elementos relevantes neste domínio: teste e validação, que fornecem parâmetros de validade e confiabilidade nos modelos gerados (c*ross validation, suplied test set, use training set, percentage split*); indicadores estatísticos para auxiliar a análise dos resultados (matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística kappa, erro médio absoluto, erro relativo médio, precisão, *F-measure*, dentre outros). E os critérios para comparar métodos e resultados de mineração de dados permitem avaliar e optar pelo melhor custo/benefício a ser adotado para a tarefa em questão. Alguns critérios relevantes neste contexto são:

- Precisão avaliativa ou preditiva: habilidade do modelo para avaliar ou prever corretamente classes, agrupamentos, regras;

- Velocidade: refere-se ao custo computacional da geração e utilização do modelo;

- Robustez: habilidade do modelo para avaliar ou prever corretamente, utilizando dados ruidosos ou com valores ausentes;

- Escalabilidade: capacidade de construir modelos eficientemente a partir de grandes volumes de dados;

- Interpretabilidade: nível de compreensão fornecido pelo modelo.

Existem diversas medidas objetivas para avaliar o grau de interesse que um padrão pode apresentar ao usuário. Tais medidas são baseadas na estrutura do padrão descoberto e em estatísticas apropriadas. Por exemplo, uma medida objetiva para avaliar o interesse de uma regra de associação é o suporte, representando a porcentagem de transações de um banco de dados de transações onde a regra se verifica. Em termos estatísticos, o suporte de uma regra $X \rightarrow Y$ é a probabilidade $P(X \cup Y)$, onde $X \cup Y$ indica que a transação contém os dois conjuntos de itens X e Y. Uma outra medida objetiva para regras de associação é a confiança,

que mede o grau de certeza de uma associação. Em termos estatísticos, trata-se simplesmente da probabilidade condicional $P(Y / X)$, isto é, a porcentagem de transações contendo os itens de X que também contém os itens de Y. Em geral, cada medida objetiva está associada a um limite mínimo de aceitação, que pode ser controlado pelo usuário. Por exemplo, o usuário pode decidir que regras cuja confiança é inferior a 0.5 devem ser descartadas como não-interessantes, pois podem simplesmente representar uma minoria ou exceção ou envolver ruídos.

Além das medidas objetivas, o usuário pode especificar medidas subjetivas para guiar o processo de descoberta, refletindo suas necessidades particulares. Por exemplo, padrões descrevendo as características dos clientes habituais de uma loja pode ser de interesse para o gerente de marketing da loja, mas com certeza é de pouco interesse para analistas que estão interessados em padrões de comportamento dos empregados da loja. Além disso, padrões que são interessantes segundo medidas objetivas podem representar conhecimento óbvio e, portanto, sem interesse. Pode-se, por exemplo, medir o grau de interesse de um padrão pelo fato de ele ser inesperado pelo usuário. Ou, ao contrário, pode-se dizer que um padrão é interessante se ele se adéqua às expectativas do usuário, servindo para confirmar uma hipótese que o usuário deseja validar.

Medidas (objetivas ou subjetivas) de avaliação do grau interesse de padrões são essenciais para a eficiência do processo de descoberta de padrões. Tais medidas podem ser usadas durante o processo de mineração ou após o processo a fim de classificar os padrões encontrados de acordo com seu interesse para um dado usuário, filtrando e eliminando os não interessantes. Em termos de eficiência é importante incorporar medidas de interesse que restrinjam o espaço de busca dos padrões durante o processo de descoberta, ao invés de após o processo ter terminado.

2. Técnicas para Classificação

Nesta seção, vamos estudar duas outras tarefas de mineração que estão de certa forma relacionadas. Trata-se das tarefas de classificação. Para cada uma destas tarefas, veremos algumas técnicas comumente utilizadas para realizá-las.

Suponha que você é gerente de uma grande loja e disponha de um banco de dados de clientes, contendo informações tais como nome, idade, renda mensal, profissão e se comprou ou não produtos eletrônicos na loja. Você está querendo enviar um material de propaganda pelo correio a seus clientes, descrevendo novos produtos eletrônicos e preços promocionais de alguns destes produtos. Para não fazer despesas inúteis você gostaria de enviar este material publicitário apenas a clientes que sejam potenciais compradores de material eletrônico. Outro ponto importante: você gostaria de, a partir do banco de dados de clientes de que dispõe no momento, desenvolver um método que lhe permita saber que tipo de atributos de um cliente o tornam um potencial comprador de produtos eletrônicos e aplicar este método no futuro, para os novos clientes que entrarão no banco de dados. Isto é, a partir do banco de dados que você tem hoje, você quer descobrir regras que classificam os clientes em duas classes : os que compram produtos eletrônicos e os que não compram. Que tipos de atributos de clientes (idade, renda mensal, profissão) influenciam na colocação de um cliente numa ou noutra classe? Uma vez tendo estas regras de classificação de clientes, você gostaria de utilizá-las no futuro para classificar novos clientes de sua loja. Por exemplo, regras que você poderia descobrir seriam: se idade está entre 30 e 40 e a renda mensal é 'Alta' então *ClasseProdEletr* = 'Sim'. Se idade está entre 60 e 70 então *ClasseProdEletr* = 'Não'.

Quando um novo cliente João, com idade de 25 anos e renda mensal 'Alta' e que tenha comprado discos, é catalogado no banco de dados, o seu classificador lhe diz que este cliente é um potencial comprador de aparelhos eletrônicos. Este cliente é colocado na classe *ClasseProdEletr* = 'Sim', mesmo que ele ainda não tenha comprado nenhum produto eletrônico.

O que é um classificador? Classificação é um processo que é realizado em duas etapas:

1. Etapa da criação do modelo de classificação. Este modelo é constituído de regras que permitem classificar as tuplas do banco de dados dentro de um número de classes pré-determinado. Este modelo é criado a partir de um banco de dados de treinamento, cujos elementos são chamados de amostras ou exemplos.

2. Etapa da verificação do modelo ou Etapa de Classificação: as regras são testadas sobre um outro banco de dados, completamente independente do banco de dados de treinamento, elas terão alta probabilidade de estarem corretas, uma vez que este banco foi usado para extraí-las. Por isso, a necessidade de um banco de dados completamente novo, chamado de banco de dados de testes. A qualidade do modelo é medida em termos da porcentagem de tuplas do banco de dados de testes que as regras do modelo conseguem classificar de forma satisfatória.

Diversas técnicas são empregadas na construção de classificadores. Neste trabalho vamos ver duas técnicas: árvores de decisão e redes neurais.

2.1 Árvore de Decisão

Uma árvore de decisão é uma estrutura que pode ser utilizada para, por meio de uma simples regras de decisão, dividir sucessivamente uma grande coleção de registros em conjuntos menores. A cada divisão realizada, os dados são separados de acordo com características em comum até chegar a pontos indivisíveis, que representam as classes.

Cada nodo da árvore representa um teste a ser realizado e as arestas definem um caminho para cada resposta desses testes. O nodo raiz é aquele que não possui nenhuma aresta de entrada, havendo zero ou mais arestas na saída. Os nodos internos possuem exatamente uma aresta de entrada e duas ou mais arestas de saída. Os nodos folhas da árvore são os pontos indivisíveis, os quais representam as classes. A Figura 2.1 ilustra um exemplo de árvore de decisão para classificar clientes como devedores ou não.

Para classificar um registro utilizando uma árvore de decisão, basta começar pelo nodo raiz da árvore, em que é aplicado o primeiro teste com o atributo referente a este nodo. O processo se repete até ser encontrado um nodo folha, o qual representa o valor associado pela árvore ao atributo classe do registro em questão.

Uma árvore de decisão é uma estrutura de árvore onde:

(1) cada nó interno é um atributo do banco de dados de amostras, diferente do atributo-classe;

(2) as folhas são valores do atributo-classe;

(3) cada ramo ligando um nó-filho a um nó-pai é etiquetado com um valor do atributo contido no nó-pai. Existem tantos ramos quantos valores possíveis para este atributo;

(4) um atributo que aparece num nó não pode aparecer em seus nós descendentes.

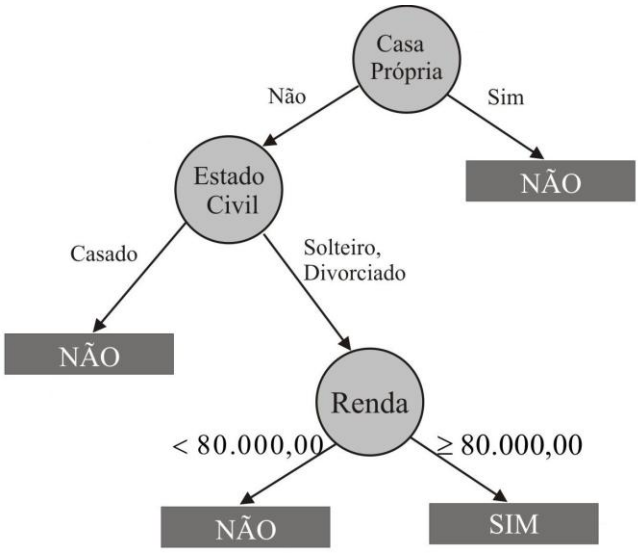

Figura 2.1: Uma árvore de decisão para classificação de clientes devedores (Ross Quinlan, 1993).

Exemplo 1 - Uma árvore de decisão é uma estrutura de árvore onde cada nó interno é um atributo do banco de dados de amostras, diferente do atributo classe, as folhas são valores do atributo-classe, cada ramo ligando um nó-filho a um nó-pai é etiquetado com um valor do atributo contido no nó-pai. Existem tantos ramos quantos valores possíveis para este atributo. Um atributo que aparece em um nó não pode aparecer em seus nós descendentes.

Considere o banco de dados de treinamento:

Nome	Idade	Renda	Profissão	ClasseProdEletr
Daniel	=< 30	Média	Estudante	Sim
João	31...50	Média-Alta	Professor	Sim
Carlos	31...50	Média-Alta	Engenheiro	Sim
Maria	31...50	Baixa	Vendedora	Não
Paulo	=< 30	Baixa	Porteiro	Não
Otávio	> 60	Média-Alta	Aposentado	Não

Tabela 2.1: Banco de dados de treinamento para árvore de decisão (Ross Quinlan, 1993).

A Figura 2.2 abaixo ilustra uma possível árvore de decisão sobre este banco de dados:

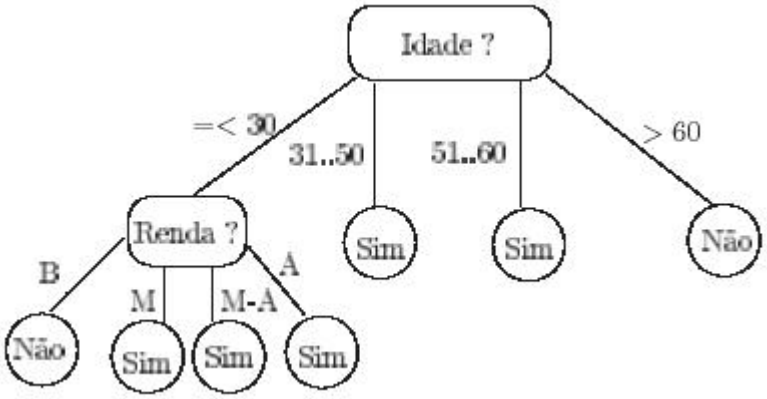

Figura 2.2: Uma árvore de decisão (Ross Quinlan, 1993).

Transformando uma árvore de decisão em regras de classificação:

Uma árvore de decisão pode ser facilmente transformada num conjunto de regras de classificação. As regras são do tipo: IF *L*1 AND *L*2 . . . AND *Ln* THEN Classe = Valor, onde *Li* são expressões do tipo Atributo = Valor. Para cada caminho, da raiz até uma folha, tem-se uma regra de classificação. Cada par (atributo,valor) neste caminho dá origem a um *Li*. Por exemplo, a árvore de decisão do exemplo acima corresponde ao seguinte conjunto de regras de classificação:

IF Idade =< 30 AND Renda = Baixa THEN Classe = Não
IF Idade =< 30 AND Renda = Média THEN Classe = Sim
IF Idade =< 30 AND Renda = Média-Alta THEN Classe = Sim
IF Idade =< 30 AND Renda = Alta THEN Classe = Sim
IF Idade 31...50 THEN Classe = Sim
IF Idade 51...60 THEN Classe = Sim
IF Idade > 60 THEN Classe = Não

Idéia geral de como criar uma árvore de decisão.

A idéia geral é a que está por trás do algoritmo ID3, criado por Ross Quinlan, da Universidade de Sydney em 1986 e de seus sucessores (um deles, o algoritmo C4.5 também proposto por Ross Quinlan em 1993).

O algoritmo C4.5 criado por Ross Quinlan (1993) trata-se do procedimento recursivo abaixo:

Gera-Arvore(*A*,Cand-List)

Input: Um banco de dados de amostras *A* onde os valores dos atributos foram categorizados, uma lista de atributos candidatos Cand-List.

Output: Uma árvore de decisão

Método

(1) Crie um nó *N*; associe a este nó o banco de dados *A*

(2) Se todas as tuplas de *A* pertencem à mesma classe C então transforme o nó *N* numa folha etiquetada por *C*. Páre.

(3) Caso contrário: se *Cand-List* = 0, então transforme N numa folha etiquetada com o valor do atributo-Classe que mais ocorre em *A*. Páre.

(4) Caso contrário: calcule Ganho(*Cand-List*). Esta função retorna o atributo com o maior ganho de informação. Será detalhada no próximo parágrafo. Chamamos este atributo de Atributo-Teste.

(5) Etiquete *N* com o nome de Atributo-Teste

(6) Etapa da partição das amostras *A*: para cada valor *si* do Atributo-Teste faça o seguinte:

(7) Crie um nó-filho *Ni*, ligado a *N* por um ramo com etiqueta igual ao valor *si* e associe a este nó o conjunto *Ai* das amostras tais que o valor de Atributo-Teste = *si*.

(8) Se *Ai* = 0 transforme o nó *Ni* numa folha etiquetada pelo valor do atributo-Classe que mais ocorre em *A*.

(9) Caso contrário: calcule Gera-Arvore(*Ai, Cand-List* - {Atributo-Teste}) e “grude” no nó *Ni* a árvore resultante deste cálculo.

Como decidir qual o melhor atributo para dividir o banco de amostras?

Agora vamos detalhar a função Ganho(*Cand-List*) que decide qual atributo em Cand-List é o mais apropriado para ser utilizado no particionamento das amostras. Para isso, utiliza-se como exemplo o banco de dados amostral da Tabela 2.2 sobre condições meteorológicas.

O objetivo é identificar quais as condições ideais para se jogar um determinado jogo.

Aparência	Temperatura	Umidade	Vento	Jogo
Sol	Quente	Alta	Falso	Não
Sol	Quente	Alta	Verdade	Não
Encoberto	Quente	Alta	Falso	Sim
Chuvoso	Agradável	Alta	Falso	Sim
Chuvoso	Frio	Normal	Falso	Sim
Chuvoso	Frio	Normal	Verdade	Não
Encoberto	Frio	Normal	Verdade	Sim
Sol	Agradável	Alta	Falso	Não
Sol	Frio	Normal	Falso	Sim
Chuvoso	Agradável	Normal	Falso	Sim
Sol	Agradável	Normal	Verdade	Sim
Encoberto	Agradável	Alta	Verdade	Sim
Encoberto	Quente	Normal	Falso	Sim
Chuvoso	Agradável	Alta	Verdade	Não

Tabela 2.2: Banco de dados amostral (Ross Quinlan, 1993).

Considera-se quatro possibilidades para a escolha do atributo que será utilizado para dividir o banco de dados no primeiro nível da árvore.

Qual a melhor escolha? Repare que se uma folha só tem 'Sim' ou só tem 'Não', ela não será mais dividida no futuro: o processo GeraArvore aplicado a esta folha pára logo no início. È desejável que isto ocorresse o mais cedo possível, pois assim a árvore produzida deverá ser menor. Assim, um critério intuitivo para a escolha do atributo que dividirá um nó seria: “Escolha aquele que produz os nós mais puros”. Por exemplo, no nosso caso, a escolha boa seria o atributo Aparência.

Grau de Pureza de um atributo num nó: Entropia. Então define-se uma função *Info* que calcula o grau de pureza de um atributo num determinado nó. Este grau de pureza representa a quantidade de informação esperada que seria necessária para especificar se uma nova instância seria classificada em 'Sim' ou 'Não', uma vez chegado a este nó. A idéia é a seguinte: se *A*1, *A*2, ..., *An* são as folhas (tabelas) saindo deste nó, *ni* = tamanho de *Ai* e *N* = total dos tamanhos das tabelas, então segundo Ross Quinlan (1993) $\text{Info(Nó)} = \sum_{i=1}^{n} \frac{n_i}{N}$ Entropia(*Ai*).

Quanto maior a entropia, maior a informação. A entropia é uma medida estatística que mede o quão “confuso” é a distribuição das tuplas entre as classes. Por exemplo, se existem 2 classes, e exatamente metade das tuplas estão numa classe e a outra metade na outra classe, então a entropia seria maximal. Por outro lado, se todas as tuplas estão numa mesma classe, então a entropia é zero.

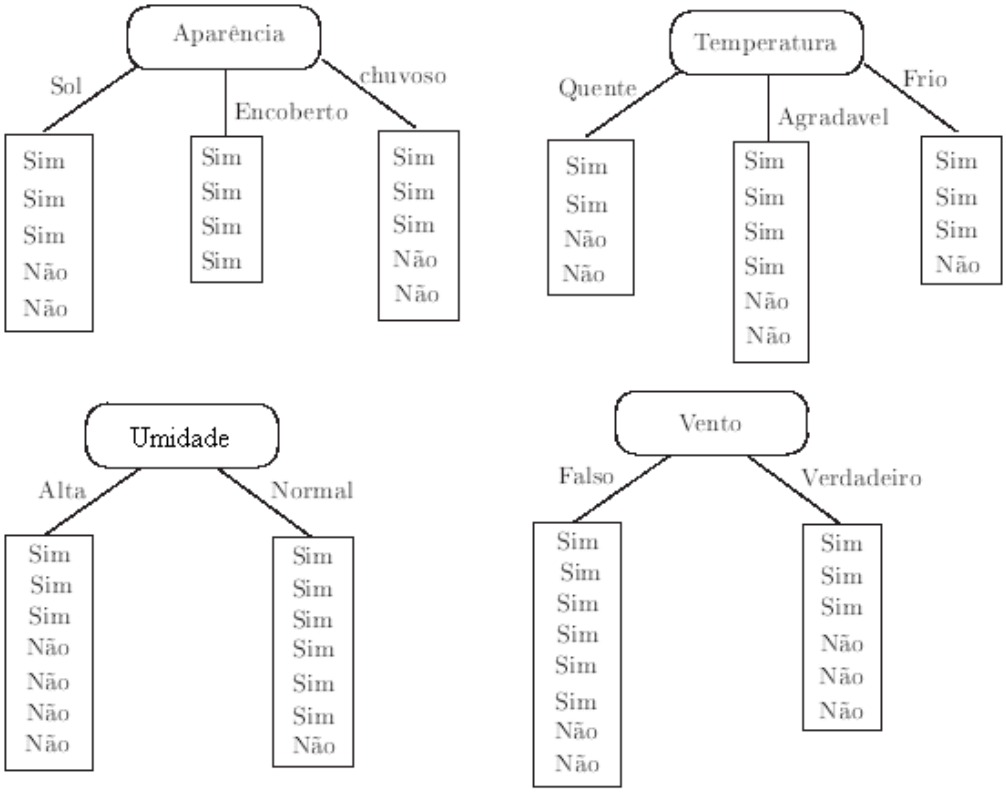

Figura 2.3: As quatro possibilidades para o atributo do nó raiz (Ross Quinlan, 1993).

Seja *Ai* uma tabela com *ni* tuplas, das quais *Si* estão classificadas como 'Sim' e *Ni* estão classificadas como 'Não'. Então a entropia de *Ai* é definida como:

Entropia (*Ai*) = $-(\frac{S_i}{n_i} log_2 \frac{S_i}{n_i} + \frac{N_i}{n_i} log_2 \frac{N_i}{n_i})$

Esta fórmula para entropia é bem conhecida e desenvolvida. Atente para o sinal negativo, necessário pois a entropia deve ser positiva e os logaritmos são negativos (já que são calculados sobre números entre 0 e 1). Esta fórmula é generalizada (da maneira óbvia) para um número de classes qualquer (Ross Quinlan, 1993).

Exemplo 2 – Considera-se quatro possibilidades para o atributo do primeiro nó.

- Se a opção for o atributo Aparência:

$$\text{Info(Nó)} = \frac{5}{14}\,\text{entropia(Folha 1)} + \frac{4}{14}\,\text{entropia(Folha 2)} + \frac{5}{14}\,\text{entropia(Folha3)}$$
$$\text{entropia(Folha 1)} = \frac{2}{5}\,log_2\frac{2}{5} + \frac{3}{5}\,log_2\frac{3}{5} = 0.971$$
$$\text{entropia(Folha 2)} = \frac{4}{4}\,log_2\frac{4}{4} + \frac{0}{4}\,log_2\frac{0}{4} = 0$$
$$\text{entropia(Folha 3)} = \frac{3}{5}\,log_2\frac{3}{5} + \frac{2}{5}\,log_2\frac{2}{5} = 0.971$$
$$\text{Logo, Info(Nó)} = \frac{5}{14}\,0.971 + \frac{4}{14}\,0 + \frac{5}{14}\,0.971 = 0.693$$

(Ross Quinlan, 1993)

- Se a opção for o atributo Temperatura:

$$\text{Info(Nó)} = \frac{4}{14}\,\text{entropia(Folha 1)} + \frac{6}{14}\,\text{entropia(Folha 2)} + \frac{4}{14}\,\text{entropia(Folha3)} = 0.911$$

(Ross Quinlan, 1993)

- Se a opção for o atributo Umidade:

$$\text{Info(Nó)} = \frac{7}{14}\,\text{entropia(Folha 1)} + \frac{7}{14}\,\text{entropia(Folha 2)} = 0.788$$

(Ross Quinlan, 1993)

- Se a opção for o atributo Vento:

$$\text{Info(Nó)} = \frac{8}{14}\,\text{entropia(Folha 1)} + \frac{6}{14}\,\text{entropia(Folha 2)} = 0.892$$

(Ross Quinlan, 1993)

Ganho de Informação ao escolher um Atributo. O ganho de informação ao escolher um atributo *A* num nó é a diferença entre a informação associada ao nó antes(*Info-pré*) da divisão e a informação associada ao nó após a divisão (*Info-pós*).

Info-pós = a informação do nó (*Info*(*Nó*)) que calculamos no passo anterior, ao escolhermos *A* como atributo divisor.

Info-pré = entropia do nó antes da divisão = $\frac{N_{Sim}}{N} log_2 \frac{N_{Sim}}{N} + \frac{N_{Nao}}{N} log_2 \frac{N_{Nao}}{N}$; onde *NSim* = total de tuplas classificadas como Sim; *NNao* = total de tuplas classificadas como Não (Ross Quinlan, 1993);

N = total de tuplas no nó.

Exemplo 3. Considera-se a situação do exemplo 2. Temos que, segundo Ross Quinlan (1993), *Info-pré* = $\frac{9}{14} log_2 \frac{9}{14} + \frac{5}{14} log_2 \frac{5}{14} = 0.940$. Logo, os ganhos de informação de cada uma das quatro escolhas são:

ganho(Aparência) = 0.940 - 0.693 = 0.247
ganho(Temperatura) = 0.940 - 0.911 = 0.029
ganho(Umidade) = 0.940 - 0.788 = 0.152
ganho(Vento) = 0.940 - 0.892 = 0.020

Logo, o atributo ideal para dividir as amostras é o atributo Aparência, como era de se supor deste o início. Veja que é o único atributo onde uma das folhas é "arrumadinha", todas as tuplas pertencendo a uma única classe.

Como transformar uma árvore de decisão em regras de classificação. Uma árvore de decisão pode ser facilmente transformada num conjunto de regras de classificação.

As regras são do tipo: IF *L*1 AND *L*2 *: : :* AND *Ln* THEN Classe = Valor, onde *Li* são expressões do tipo Atributo = Valor. Para cada caminho, da raiz até uma folha, tem-se uma regra de classificação. Cada par (atributo,valor) neste caminho dá origem a um *Li*.

Por exemplo, a árvore de decisão do exemplo 1.1 corresponde ao seguinte conjunto de regras de classificação.

- IF Idade =< 30 AND Renda = Baixa THEN Classe = Não
- IF Idade =< 30 AND Renda = Média THEN Classe = Sim
- IF Idade =< 30 AND Renda = Média-Alta THEN Classe = Sim
- IF Idade =< 30 AND Renda = Alta THEN Classe = Sim
- IF Idade 31..50 THEN Classe = Sim
- IF Idade 51..60 THEN Classe = Sim
- IF Idade > 60 THEN Classe = Não

2.1.1 Crescimento e poda

As árvores de decisão são construídas usando um algoritmo de partição recursiva. Cada divisão binária de um nó *t* gera dois nós descendentes denotados por *tY* e *tN* se a resposta for "Sim" ou "Não" a uma única pergunta adotada para ao nó *t*. Cada divisão deve gerar subconjuntos que possuem classes cada vez mais homogêneas em comparação com o subconjunto do antepassado. Se ocorrer uma divisão de um nó para que os pontos que pertencem às classes *w1*, *w2* formem o subconjunto X*tY*, e os pontos das classes *w3*, *w4* formam o subconjunto X*tN*, então os novos subconjuntos são mais homogêneos em comparação a X*t* ou "puros" na terminologia de árvore de decisão. O objetivo, então, é definir uma medida que quantifica os erros (impureza) no nó e a divisão no nó para que a impureza total nos nós descendentes seja reduzida em relação à impureza no nó antepassado (Theodoridis et al., 2006; Safavian et al., 1991).

A probabilidade de que um vetor no subconjunto X*t*, associado a um nó *t*, pertença à classe *wi*, *i* = 1, 2, ..., M, é denotada por P (*wi*/*t*). Uma definição comumente usada de impureza de nó, é denotada como I(t), é dada por (Theodoridis et al., 2006; Safavian et al., 1991):

$$I(t) = -\sum_{i=1}^{M} P(w_i / t) \log_2 P(w_i / t)$$

Isto não é nada mais do que a entropia associada ao subconjunto X*t*. Na prática, a probabilidade é prevista pelas respectivas porcentagens, N*it*/N*t*, onde Nt é o número de pontos em X*t* e N*it* é o número de pontos em X*t* que pertencem à classe *wi*. Assumindo que executando uma divisão, os pontos de N*tY* são enviados como "Sim" ao nó (X*tY*) e os pontos N*tN* como "Não" ao nó (X*tN*). A redução na impureza de nó, DI(*t*), é definida como:

$$\Delta I(t) = I(t) - \frac{N_{tY}}{N_t} I(t_Y) - \frac{N_{tN}}{N_t} I(t_N)$$

Onde *I*(*tY*), *I*(*tN*) são as impurezas dos nós *tY* e *tN* respectivamente. O objetivo agora é adotar do conjunto de candidato às perguntas, aquela que executa a divisão que leva à mais alta redução da impureza.

O critério de poda define quando cada nó deve deixar de se dividir e passa a ser uma folha da árvore. Uma vez que se declara que um nó é uma folha, então tem de ser dado um rótulo às amostras que atingirem esta folha. Uma regra comumente usada é a regra de maioria, isto é, a folha é rotulada como *wj* quando a maioria dos vetores X*t* pertencem a classe w*j*. Após a construção de uma árvore de decisão é importante avaliá-la. Esta avaliação é realizada pela utilização de dados que não tenham sido usados no treinamento. Esta estratégia permite estimar como a árvore generaliza os dados e se adapta a novas situações, podendo, também, estimar a proporção de erros e acertos ocorridos na construção da árvore (Theodoridis, 2006; Safavian, 1991).

2.1.2 Vantagens obtidas com o uso das árvores de decisão

Dentre as técnicas de mineração de dados, o método com uso de árvores de decisão tem várias vantagens:

- São fáceis de entender e interpretar. Pessoas conseguem entender modelos em árvore depois de uma simples explicação, ajudando na documentação e entendimento do analista com o usuário.
- Preparação dos dados para uma árvore de decisão é dispensável ou desnecessário. Outras técnicas normalmente requerem normalização dos dados, variáveis de teste precisam ser criadas e campos em branco precisam ser removidos.
- Está apto a lidar tanto com dados nominais ou com dados categóricos.
- Outras técnicas são normalmente especializadas em analisar conjuntos de dados que têm somente um tipo de variável. Ex. regras de relacionamento só podem ser usadas com variáveis nominais enquanto redes neurais podem ser usadas somente com variáveis numéricas.
- É um modelo "caixa-branca". Se uma dada situação é observada em um modelo, a explicação é facilmente obtida pela de lógica booleana. Um exemplo de um modelo "caixa-preta" é uma rede neural já que a explicação para os seus resultados obtidos é excessivamente complexa de compreender.
- É possível validar um modelo usando testes estatísticos. Torna fácil verificar a confiabilidade do modelo, ajudando a encontrar situações que não foram totalmente especificadas e possibilitando no fornecimento de maneiras mais objetivas de identificar todas as combinações possíveis.
- É robusto, e tem bom desempenho com grandes quantidades de informação em pouco tempo.
- É um Método não-paramétrico:
  - Não assume nenhuma distribuição particular para os dados.
  - Pode construir modelos para qualquer função desde que o numero de exemplos de treino seja suficiente.
- A estrutura da arvore de decisão é independente da escala das variáveis:
  - Transformações monótonas das variáveis (log x, 2*x, ...) não alteram a estrutura da arvore.

- Elevado grau de interpretabilidade:
  - Uma decisão complexa (prever o valor da classe) é decomposto numa sucessão de decisões elementares.
- É eficiente na construção de modelos:
  - Complexidade média O(n log n)
  - Robusto á presença de pontos extremos e atributos redundantes ou irrelevantes.

2.1.3 Algumas das desvantagens obtidas com o uso das árvores de decisão

Apesar das inúmeras vantagens apresentadas referentes a técnica de Árvore de Decisões, existem algumas desvantagens que devem ser lembradas neste trabalho e dentre algumas das desvantagens estão:

- A dificuldade de saber em alguns casos quando iniciar uma formulação na árvore de decisão e também a dificuldade do entendimento de um determinado conceito representado como uma árvore de decisão grande.
- Os muitos dos usuários que não estarão familiarizados com a tabela de decisões.
- Ela não nos fornece um quadro nítido da estrutura.
- As árvores de decisão são complexas e de difícil entendimento para aqueles que nunca tenham-na visto antes.
- A limitação a uma linguagem descritiva baseada em atributos-valores.
- Um mesmo conceito pode ser representado por várias árvores de decisão.

3. Conclusão

No breve espaço deste trabalho, minha principal missão foi introduzir um pouco do pensamento que está por trás da técnica de Data Mining. Obviamente, ainda há muito a se falar sobre o assunto (*clustering*, redes neurais, métodos genéticos, mineração em textos, *roll up/drill down*, etc), mas é importante notar que em praticamente todos esses casos o que se deseja é descobrir padrões em volumes de dados. É importante ressaltar também que o Data Mining não é o final da atividade de descoberta de conhecimentos, mas é tão somente o início. É imprescindível (ao menos com a tecnologia atual) dispor de analistas capacitados que saibam interagir com os sistemas de forma a conduzi-los para uma extração de padrões úteis e relevantes.

A árvore de decisão é muito útil como uma técnica exploratória. Contudo, ela não tenta substituir métodos estatísticos tradicionais existentes. Muitas outras técnicas podem ser usadas para classificar ou predizer um grupo de ocorrências a um conjunto predefinido de classes como redes neurais artificiais, máquinas de vetores suporte, entre outras.

O método de árvore de decisão reconhece que existem dois fatores importantes que afetam o futuro – escolha e probabilidade. E ao avaliá-los teremos que considerar dois parâmetros – custos e conseqüências.

Ao construir uma árvore de decisão, é possível fazermos uma análise para determinar a escolha mais favorável, levando em consideração os custos, as probabilidades e as conseqüências associados. Primeiro, devemos seguir pela árvore de decisão em cada caminho lógico alternativo e o seu valor se calcula acumulando os custos e benefícios do começo ao fim. Então, utilizando estes valores e verificando o retorno de cada caminho lógico alternativo, o "valor esperado" de cada escolha é calculado, levando em conta as conseqüências quando as probabilidades ocorrerem.

Existem vários desafios de utilizar as árvores de decisão de maneira eficaz, incluindo a limitação prática do método de analisar um número pequeno de opções de decisão com uma gama limitada de riscos possíveis. O projeto típico envolve muitas decisões em níveis diferentes, cada uma com uma ampla gama de riscos associados, e tentar demonstrar isso somente em uma árvore de decisão pode resultar em um modelo enorme e sem utilidade. A técnica requer também que todos os fatores sejam representados de forma quantitativa – custos e conseqüências se expressam normalmente em termos financeiros, e a probabilidade deve ser estimada para todas as possibilidades. E a árvore de decisão supõe também a

existência de “uma pessoa que toma decisão neutra aos riscos”, cujas escolhas se baseiam em valor esperado mais alto – o que acontece raramente. Apesar dessas limitações, a análise de árvore de decisão apresenta uma técnica quantitativa poderosa para avaliar futuros possíveis, levando em consideração os efeitos tanto da escolha como da probabilidade e estimando ambos os custos e conseqüências.

Apesar do aumento do uso de modelos de árvore de decisão, ainda existem muitos campos na área de Tecnologia da Informação, principalmente a área de Inteligência de Negócio (*Business Intelligence*) que podem se beneficiar desta técnica. Portanto, o potencial do uso de árvore de decisão ainda pode ser muito explorado em vários campos das áreas de pesquisa e desenvolvimento, ainda são poucas as empresas que conhecem e se utilizam destas técnicas. Nas áreas de marketing de bancos e empresas de telecomunicações existem vários casos de sucesso com a aplicação de Data Mining. As empresas que já utilizam, podem descobrir fatos que representam significativas vantagens, e podendo estar sempre um passo a frente sobre seus concorrentes.

E um dos exemplos mais divulgados é o da cadeia americana *Wal-Mart*, que identificou um hábito curioso dos consumidores. Há cinco anos, ao procurar eventuais relações entre o volume de vendas e os dias da semana, o *software* de Data mining utilizado, apontou que às sextas-feiras, as vendas de cervejas cresciam na mesma proporção que as de fraldas, e uma investigação mais detalhada, com a ajuda deste software de Data minig, revelaram que ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cerveja para o final de semana, e com essas eventuais informações em mão a rede de hipermercados *Wal-Mart* iniciou inúmeras promoções referentes aos produtos, cervejas e fraldas. Em vista do exemplo podemos entender a importância da utilização da Data mining para avaliar em quais situações a empresa pode por em pratica determinadas ações, e tomar determinadas decisões para melhoria das vendas e consequentemente aumento do lucro.

Há quem consiga aumentar a receita da empresa, cortar gastos ou até detectar fraudes, como é o caso do *Bank of America*, um dos maiores bancos americanos situado no estado da Carolina do Norte nos Estados Unidos da América, usou as técnicas citadas acima para selecionar entre seus 36 milhões de clientes aqueles com menores riscos de dar calote num empréstimo. A partir desses relatórios, enviou cartas oferecendo linhas de crédito para os correntistas cujos filhos tivessem entre 18 e 21 anos e, portanto, precisassem de dinheiro para ajudar os filhos a comprar o próprio carro, uma casa ou arcar com os gastos da faculdade. E como resultado, em três anos, o banco lucrou 30 milhões de dólares.

Em Inteligência de Negócio, a Data Mining poderá trazer inúmeras vantagens aos profissionais através do acesso às informações sobre o comportamento do cliente no e-commerce, como a identificação de padrões de consumo dos clientes. Possibilitará uma ação rápida e direcionada a perfis individualizados de clientes. Entre outras análises, será possível, por exemplo, prever o aumento ou redução nas vendas de um determinado produto em virtude da variação dos seus preços, determinarem o preço ótimo de forma a maximizar o lucro da empresa, como visto no exemplo da rede *Wal-Mart*, determinar a influência nas vendas de uma atividade de marketing, conhecer a influência da variação de preço dos concorrentes nas suas vendas e conhecer a influência da variação do seu preço nas vendas dos concorrentes. Também será possível obter conhecimento das atividades de marketing dos concorrentes e sua influência nas vendas, segmentar o mercado em função do comportamento de seus clientes, de modo a possibilitar campanhas de marketing específicas para cada segmento e extrair conhecimento da estratégia dos concorrentes na formação de preço.

4. Referências Bibliográficas

DINIZ, C.A. & LOUZADA NETO, F. **Data Mining: uma introdução**. São Carlos: Associação Brasileira de Estatística, 2000.

MORETTIN, P.A. & TOLOI, C.M. **Séries Temporais. 2.ª ed. São Paulo: Atual**, 1987.

MATTAR, F.N. **Pesquisa de Marketing**. São Paulo: Atlas, 1998.

GUJARATI, D.N. **Econometria Básica. Trad. Ernesto Yoshita. São Paulo**, 2000.

PADOVANI, C.R. **Estatística na Metodologia da Investigação Científica.** Botucatu: UNESP, 1995.

GOLDSCHIMIDT, RONALDO; Passos, Emmanuel. **Data Mining: Um guia Prático. Rio de Janeiro**, 2005.

SAFAVIAN, S. R.; Landgrebe, D. **A Survey of Decision Wee Classifier Methodology. IEEE Transactions on systems, man, and cybernetics**, 1991.

THEODORIDIS, S.; Koutroumbas, K. **Pattern recognition Academic Press**, 2006.

D. HAND, H. Mannila, P. Smith. **Principles of Data Mining. MIT Press**, 1998.

FAYYAD, Usama; Piatetski-Shapiro, Gregory; Smyth, Padhraic. **The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM**, 1996