DATA MINING NO CONTEXTO DE *CUSTOMER RELATIONSHIP MANAGEMENT*

TUTORIAL – MÉTODOS QUANTITATIVOS E INFORMÁTICA

Fernando Carvalho de Almeida
Professor Doutor da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo – FEA/USP
E-mail: fernando_c_de_almeida@yahoo.com.br

José de Oliveira Siqueira
Professor Doutor da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo – FEA/USP
E-mail: siqueira@usp.br

Luciana M. Onusic
Mestranda da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo – FEA/USP
E-mail: luciana_onusic@yahoo.com

Recebido em: 24/05/2004

Aprovado em: 02/08/2005

RESUMO

As empresas têm dado importância crescente ao conhecimento do cliente, de maneira a melhorar o relacionamento com ele e aumentar a sua fidelidade com elas. Os processos de gerenciamento do relacionamento com o cliente, ou CRM (*customer relationship management*), integram um processo de aquisição de conhecimento sobre o perfil e o comportamento dos clientes chamado de *data mining* (DM). O objetivo principal deste artigo é apresentar as etapas do processo de DM que devem orientar a sua implementação no contexto do CRM. Algumas técnicas estatísticas freqüentemente utilizadas em DM são apresentadas.

Palavras-chave: Aquisição de conhecimento, Gerenciamento do relacionamento com o cliente (CRM), Garimpagem de dados, *Marketing* de relacionamento.

DATA MINING IN THE CONTEXT OF CUSTOMER RELATIONSHIP MANAGEMENT

ABSTRACT

Companies are increasingly concerned about understanding customer behavior in order to improve the relationship and loyalty of customers towards the company. Customer Relationship Management CRM therefore integrates an information acquisition process called Data Mining DM to acquire knowledge about the profile and behavior involved. The sequence of DM is presented together with several statistical techniques to orient the implementation of DM in the context of CRM.

Key words*: Knowledge acquisition, Customer relationship management CRM, Data mining DM, Relationship marketing.*

1. INTRODUÇÃO

Estima-se que a cada 20 meses as empresas no mundo dobrem o volume de dados acumulados em seus computadores (WITTEN e FRANK, 2000). Em particular, o uso de sistemas de informação no relacionamento das empresas com seus clientes as tem feito acumular grande quantidade de dados sobre eles. A partir do conhecimento contido nesses dados, uma empresa pode conhecer e entender melhor seus clientes, e oferecer produtos mais direcionados às suas demandas. Esses dados podem auxiliar também na distinção entre clientes adequados e inadequados para o negócio da empresa. Isso é possível por meio de um processo organizado de transformação de dados em informação e, em seguida, em conhecimento. O dado pode ser definido como um número com um contexto. A informação é um dado com significado, cuja vida útil pode ser instantânea para o decisor. Um conhecimento é uma informação com vida útil não instantânea. Esse processo é chamado de *data mining* (DM) e incorpora ferramentas de Tecnologia de Informática (TI), conhecimento de gestão de dados e análise estatística exploratória de dados multivariados. DM, do ponto de vista estatístico, é a análise estatística exploratória multivariada de grandes quantidades de dados numéricos.

Os dados para DM estão dentro e fora da empresa. As transações operacionais efetuadas por meio do uso de sistemas de informação geram esses dados, cujo conhecimento implicitamente acumulado pode ser formalizado, ou estruturado, por meio de técnicas tais como redes neurais artificiais, árvores de classificação e regressão, regras de associação, entre outras (DE ALMEIDA, 1995; HAIR JR. *et al.*, 1998; BERRY e LINOFF, 2000; WITTEN e FRANK, 2000; etc.). Dados e bases de dados externos, acessados, entre outros meios, através da internet, constituem também uma fonte de dados (WEIR, 2000). Todavia, da obtenção dos dados à disponibilização e utilização do conhecimento gerado, um processo complexo é desenvolvido, envolvendo diversos fatores que, se forem mal dimensionados ou desprezados, podem levar ao fracasso dessas iniciativas bem intencionadas.

O DM tem gerado interesse, particularmente como apoio ao gerenciamento do relacionamento com os clientes de uma empresa, ou CRM (*customer relationship management*). O processo de CRM associa conceitos de *marketing* de relacionamento ao uso de TI (BERRY e LINOFF, 2000). CRM pode ser definido como gerenciamento contínuo do relacionamento com o cliente, garantindo melhor atendimento, fidelização e novas oportunidades de negócios. O objetivo do CRM é tornar a empresa capaz de tratar consistentemente suas interações com os clientes por meio de canais e funções, e assim construir, manter e aperfeiçoar as relações com eles. Há alguns fatores que aumentam a complexidade do relacionamento com o cliente e justificam a adoção do CRM: (i) compressão do tempo entre a necessidade e a sua satisfação, (ii) competição acirrada em preço e aumento de custos operacionais, (iii) consumidores mais exigentes, que provocam o aumento das variantes dos serviços e das maneiras de oferecê-los, e (iv) competição em nichos. A interação entre *strategic business intelligence* (SBI) e DM propicia à empresa (i) a manutenção do sucesso de mercado (sobrevivência sustentada), (ii) o aprendizado com os relacionamentos e com os dados gerados a partir deles (processo cumulativo de conhecimento do cliente) e (iii) o foco no cliente e não no produto/serviço (visão unificada e integrada do cliente).

O processo de DM bem-sucedido deve levar a organização a imprimir em seus serviços e produtos as seguintes características fundamentais: (i) serviço personalizado (Eu sou alguém! Sou um ser humano diferente dos outros!), (ii) atendimento pessoal (Já sou da casa!), (iii) antecipação da necessidade (Era isso o que eu estava precisando! Parece que adivinharam o que eu necessitava!) e cobertura completa das necessidades num tema (Todos os produtos/serviços complementares estão disponíveis!). A competição da década de 90 do século passado caracterizou-se pelo aumento da eficiência e da velocidade computacionais. Já a do século XXI consiste numa competição baseada em modelos de negócio e na habilidade de adquirir, acumular e usar eficazmente o conhecimento gerado nas organizações. Portanto, automação e eficiência não são suficientes para garantir o sucesso do negócio. Flexibilidade e prontidão nas respostas às mudanças ambientais provocadas pela economia digital são as características que as organizações devem adquirir em face do darwinismo que impera no mundo dos negócios. Há, no entanto, algumas barreiras no mercado brasileiro para a implantação do *marketing* de relacionamento, oriundas da infra-

estrutura socioeconômica, política, legal e cultural, explicada pela insuficiência de formação educacional da população. Entretanto, no caráter brasileiro há a vocação para o bom atendimento do cliente. A gestão do conhecimento (*knowledge management* – KM) é uma disciplina que promove uma abordagem integrada que sustente a criação, retenção, acesso, compartilhamento e alavancagem dos ativos intangíveis da organização, para que haja ganho no negócio. O conhecimento é entendido como um redutor de incertezas que depende do contexto, alternativas, fatos, dados, informações, experiências, especialidades individuais e coletivas, etc. A KM é fundamental para o processo de tomada de decisão. Além disso, conforme Francis Bacon, conhecimento é poder.

O presente artigo é desenvolvido a partir de um estudo de caso em uma empresa de grande porte do setor financeiro e possui dois objetivos. Num primeiro momento, explora as etapas do processo de *data mining* na empresa, identificando e explorando as características de cada uma delas. Essa discussão é voltada particularmente ao processo de gestão do relacionamento com o cliente (CRM). Sendo assim, são discutidas as dimensões do processo e o estabelecimento de uma estratégia de DM. A partir de atividades de DM realizadas pelos autores na empresa objeto deste estudo, foi possível identificar os processos que são apresentados neste texto. Por questões de confidencialidade, o nome da empresa não é revelado. Em seguida, cumpre-se o segundo objetivo do trabalho, no capítulo "Tecnologias para DM", no qual são discutidas algumas técnicas básicas passíveis de serem aplicadas em DM. Procura-se, nesse capítulo, organizar as técnicas segundo sua forma de aquisição do conhecimento.

O trabalho traz, então, tanto contribuições obtidas em campo, a partir da aplicação de técnicas de DM em uma determinada empresa, quanto conclusões a respeito dos tipos de ferramentas aplicáveis ao processo de DM.

Sendo assim, o trabalho está organizado em duas partes. Em uma primeira parte, nos itens 2 e 3, destacam-se os conceitos de DM e de CRM, bem como a inter-relação das duas atividades. A partir do item 4 são discutidas as dimensões do processo de DM, com base nas atividades realizadas em campo pelos autores. Propõe-se, dessa maneira, um modelo original de representação do processo de DM na empresa.

2. OS OBJETIVOS DO PROCESSO DE DM

O objetivo dos processos de DM é a captura de padrões subjacentes às grandes bases de dados (vide DUDA, HART e STORK, 2001). Um termo também associado à DM é aquisição de conhecimento (*knowledge discovery* - KD). Os processos chamados DM têm como objetivo garimpar dados com potencial conteúdo informacional para posterior construção de conhecimento. A KD possui como objetivo a descoberta de conhecimento. Procura-se identificar, por exemplo, padrões de comportamento de clientes atuais ou de clientes-alvo.

A DM pode ser definida mais precisamente como o processo de reconhecimento, extração e acompanhamento de padrões e regras de produção latentes e potencialmente úteis, relativos ao comportamento dos clientes, a partir de grandes bases de dados organizacionais; seu objetivo é criar modelos para a tomada de decisão, com o intuito de prever o comportamento dos clientes, baseados em suas atividades registradas (vide DUDA *et al.*, 2001; SCHÜRMANN, 1996; ROSENFELD e WECHSLER, 2000). É importante notar que os dados analisados são de natureza observacional. A DM pode ser utilizada para extração de conhecimento, visualização e/ou correção de dados. Do ponto de vista organizacional, a DM pode ser aplicada para retenção de cliente, campanha (promoção para aquisição), serviço de atendimento e vendas, detecção de fraude e avaliação de risco, etc. Ela consiste num processo de aquisição de conhecimento e, portanto, não se resume a um sistema informatizado. Para sua implantação é necessária a formação de uma equipe multidisciplinar, que tenha o espírito de começar pequena e avançar sempre.

Conforme BERRY e LINOFF (2000), a analogia de DM com a arte da fotografia é útil. A empresa certa vez lançou o seguinte *slogan* para suas câmeras: "*You press the button, we do the rest*." Esse *slogan* resume a maior tentação a que uma organização está sujeita, a saber, a de concluir que a simples compra de um *software* de DM pode ser suficiente para extrair conhecimento e pô-lo em prática com sucesso. A DM pode ser adotada pela organização numa das quatro seguintes abordagens: (i) comprar uma câmera Polaroid: comprar escores, (ii) comprar uma câmera completamente automática: comprar *softwares* especializados numa

determinada aplicação que embutem ferramentas de DM (*e.g.*, detecção de fraude), (iii) contratar um fotógrafo de casamento: contratar consultores externos para construir os modelos para projetos específicos e/ou (iv) tornar-se um fotógrafo profissional: dominar a arte de DM dentro da própria organização.

O aumento do poder de tratamento e da capacidade de armazenamento de dados dos computadores, aliado à facilidade de uso e poder dos *softwares* de DM, tem permitido explorar intensamente bases de dados geradas a partir das transações operacionais das empresas. São diversas as aplicações de DM em Administração. As suas primeiras aplicações foram feitas em Finanças. Uma vez que para fazer DM é preciso dados, a área financeira pôde fornecê-los em abundância. As aplicações mais comuns foram em avaliação de risco de crédito, seguros, insolvência, hipotecas, etc. (DE ALMEIDA, 1995; WONG, BODNOVICH e SELVI, 1997).

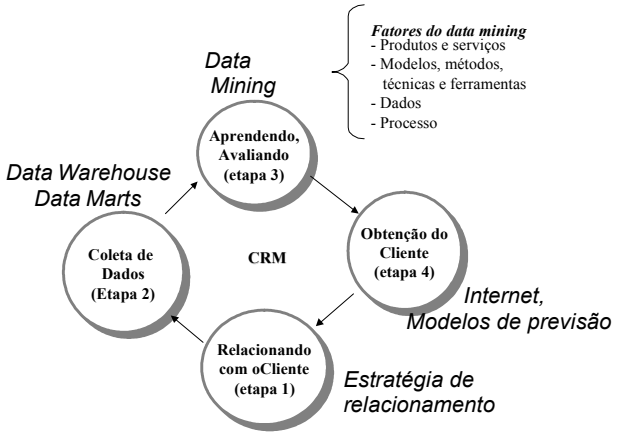
3. DM NO CONTEXTO DE CRM

Mais recentemente tem havido um significativo foco da DM na gestão do relacionamento com o cliente (CRM – *Customer Relationship Management*) (BERSON, SMITH e THEARLING, 2000). A partir das possibilidades da TI e dos conceitos de *marketing* de relacionamento, as técnicas de DM contribuem para o sucesso de uma relação eficaz e de lealdade do cliente para com a empresa. Clientes leais são mais rentáveis, geram menor custo e maior valor para a empresa (BRETZKE, 2000; REICHHELD, 1996; VAVRA, 1995). Do *marketing* direto ao *marketing* 1:1, a idéia de passar de uma estratégia de *marketing* de massa a uma estratégia de comunicação e de relação individualizada cliente-empresa, focando os anseios de cada cliente e não de uma massa de clientes, foi popularizada por Don Peppers (PEPPERS, ROGERS e DORF, 1999). A TI, em seu atual estágio de desenvolvimento, traz a oportunidade de que cada cliente seja tratado de maneira personalizada, como se fosse parte de um pequeno grupo de clientes, ainda que na realidade possa ser parte de uma carteira de milhões de clientes.

Segundo PEPPERS, ROGERS e DORF (1999), o segredo de uma boa relação com o cliente é memória: uma relação bem-sucedida com o cliente baseia-se em se lembrar de pequenos detalhes. *Marketing* de relacionamento, *marketing* 1:1 ou ainda CRM significam aprender e entender o que cada cliente espera da empresa e assim poder suprir seus anseios. Pode-se dizer que as tecnologias de DM buscam desenvolver, manter e usar a memória sobre o cliente, seu comportamento e seus interesses. Memória não significa apenas armazenar dados sobre o cliente, mas armazenar conhecimento sobre o cliente. O processo de DM visa tornar esse conhecimento explícito, operacionalizável e aplicável.

A DM é uma das principais atividades que objetivam extrair conhecimento a partir dos dados gerados pelo relacionamento com o cliente. Ela pode ser entendida como um dos quatro elementos do ciclo do CRM representados na Figura 1: (1) relacionamento com o cliente, (2) coleta de dados gerados a partir da relação com o cliente e armazenamento dos dados em bases de dados (*data warehouse* e *data mart*), (3) DM e (4) obtenção dos clientes.

Figura 1: Ciclo de CRM



Fonte: Elaborada pelo autor.

Um dos pontos de entrada do ciclo do CRM é a obtenção dos clientes, que é fruto de uma estratégia de relacionamento criada pela empresa (etapa 4). A partir da obtenção dos clientes, inicia-se um relacionamento empresa-cliente (etapa 1). Por meio da utilização de recursos de TI, a empresa poderá captar dados sobre seus clientes e armazená-los em bases de dados gerenciais (*data warehouses* e *data marts* - etapa 2). O passo final é a exploração desses dados, que permite aprender com a relação desenvolvida na etapa 3, com o intuito de aumentar

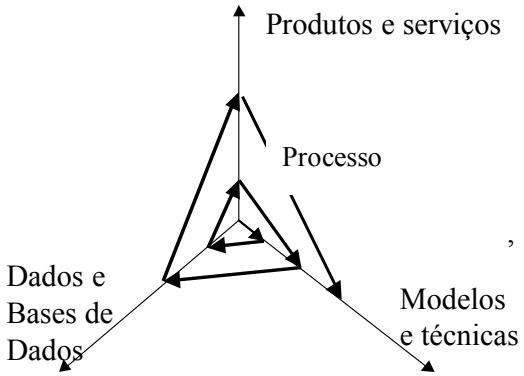
o conhecimento sobre o cliente. Há retroalimentação no processo, pois, ao coletar dados sobre os clientes (etapa 3), a empresa procurará aprender e conhecer o cliente a partir do processo de DM (etapa 4). Isso irá incrementar a eficácia de seu processo de obtenção do cliente (etapa 4) e de sua estratégia de relacionamento (etapa 1), levando à obtenção de novos dados (novamente a etapa 3) com uma nova estratégia de relacionamento. Algumas aplicações de DM voltadas à análise do relacionamento com o cliente são: análise do perfil e características do cliente, planejamento de esforços de *marketing* para um novo produto, desempenho futuro de um novo produto, definição do preço de novos produtos e serviços, administração do atrito com o cliente, estimação do valor do cliente, propensão dos clientes à compra de produtos, análise da relação entre produtos e venda de produtos casados. Uma descrição mais detalhada das aplicações de DM pode ser encontrada em BERRY e LINOFF (2000).

4. DIMENSÕES DO PROCESSO E ESTABELECIMENTO DE UMA ESTRATÉGIA DE DM

Uma vez que uma organização decide implementar um processo de DM, ela se vê diante de um conjunto de alternativas e caminhos possíveis para desenvolver o processo. Há três dimensões a serem consideradas para que haja sucesso na adoção da DM. Essas dimensões são as componentes principais com que a organização deverá lidar, independentemente da estratégia que for adotar, e estão representadas na Figura 2: (1) dados e bases de dados, isto é, a fonte do conhecimento a ser explorado; (2) produtos e serviços sobre os quais pode desenvolver o processo de DM; e (3) modelos conceituais e técnicas de DM.

Cada uma dessas dimensões representa um conjunto de vias que se ramificam, oferecendo à empresa possibilidades de escolha e dificuldades na decisão sobre qual caminho tomar. A seguir serão abordadas todas essas dimensões e destacados alguns problemas e suas possíveis soluções.

Figura 2: Três dimensões da DM



Fonte: Elaborada pelo autor.

4.1. Dimensão produtos e serviços

Uma seguradora, por exemplo, que deseja implementar um processo de DM, possui uma carteira grande de produtos/serviços que podem ser oferecidos a um cliente. Por qual produto ou serviço deve-se iniciar a DM? Por todos os produtos/serviços simultaneamente? Por um que a empresa já conhece bem e sobre o qual tem bastante experiência, ou por um produto/serviço recém-lançado? Se a empresa já conhece bem o produto, a chance de já possuir um bom volume de dados sobre ele é maior. No entanto, o produto bem conhecido pode ser maduro, com menores oportunidades no mercado, ao contrário de um novo produto, para o qual o mercado potencial ainda é grande. A respeito de um novo produto/serviço, entretanto, a empresa pode ter poucos dados. Delineia-se aqui uma questão de oportunidade, a ser avaliada pela empresa para a escolha do produto pelo qual começará o processo de DM.

4.2. Dimensão modelos e técnicas

A segunda dimensão decisória da DM é a escolha das ferramentas e técnicas para explorar os dados gerados pelo relacionamento com o cliente, o que permitirá à empresa conhecê-lo, identificar padrões de comportamento, tipos de cliente, e também tomar decisões com o auxílio de previsões geradas por esses modelos (*e.g.*, decisão de envio de mala direta a partir de um modelo de propensão à compra). Há várias técnicas e modelos que podem ser aplicados nesta etapa do processo. Qual modelo

e técnica poderiam ser utilizados para um dos seus problemas? A empresa quer aprofundar o conhecimento do perfil de seus clientes, procurando segmentá-los em subgrupos homogêneos. Uma empresa orientada para o cliente costuma usar técnicas estatísticas convencionais para criar modelos de segmentação de mercado e análise de perfil do cliente. Duas questões surgem: (i) Como as novas técnicas de DM poderiam melhorar a capacidade preditiva de seus modelos (redes neurais artificiais, árvore de classificação e regressão, etc.)? e (ii) Se as técnicas convencionais não estão produzindo resultados interessantes, por quais técnicas podem ser substituídas? A resposta para a segunda questão pode ser a seguinte: em geral, a cada nova campanha o modelo perde capacidade preditiva. Além disso, o problema de desempenho de modelos de previsão pode estar ligado à maturidade dos produtos/serviços. Num mercado saturado o modelo de previsão identifica com facilidade um cliente potencial, porém ele provavelmente já possui o produto/serviço. Então, a solução do problema decorrente de um mercado saturado possui dois aspectos: (i) é preciso criar uma estratégia de inovação do produto/serviço; (ii) todo modelo de previsão deve levar em conta o ciclo de vida do produto no mercado, a fim de identificar a probabilidade de o cliente potencial já ter produto similar de um concorrente.

4.3. Dimensão dados e base de dados

Para a obtenção de conhecimento sobre o cliente são necessários dados. Estes são coletados a partir do relacionamento com o cliente. Quando numa empresa um volume expressivo de dados foi acumulado em base de dados informatizada, pode-se dar início ao processo de DM. No entanto, a utilidade e disponibilidade desses dados devem ser analisadas, fazendo com que haja um esforço específico a ser realizado nesta dimensão do processo. As decisões sobre por onde começar a análise dos dados, as técnicas a utilizar e os dados a analisar estão intimamente relacionadas.

5. O PROCESSO DE DM

O último aspecto da DM é o seu processo em si. Não é viável, na prática, caminhar e esgotar o trabalho numa das dimensões de cada vez. Isto é, não é possível, em primeiro lugar, criar uma estratégia rígida, por meio da qual se delineariam todas as ações necessárias à implementação do DM, para depois passar às outras dimensões. Não é viável atacar a dimensão de dados e procurar resolver todas as questões referentes à disponibilidade de dados antes de conhecer mais o problema que se quer resolver. Vale ressaltar que esta é uma tentação muito grande das empresas, que as compele a fazer um grande trabalho de criação do *data warehouse* da empresa (algumas empresas podem ter um excessivo gasto na construção de um *data warehouse* corporativo, sem saber a finalidade desse empreendimento), alimentando-o de maneira pouco refletida com todos os dados contidos nas bases de dados operacionais. Fazer isso significaria perder o foco do problema e pensar que, uma vez reunido um grande volume de dados, a consulta gerencial far-se-ia como conseqüência. Procedendo-se dessa forma, tem-se a ilusão de que os dados organizados de maneira acessível, sem saber-se exatamente para que fim e como serão usados, são suficientes para fazer com que os executivos se concentrem neles à procura de respostas para seus problemas.

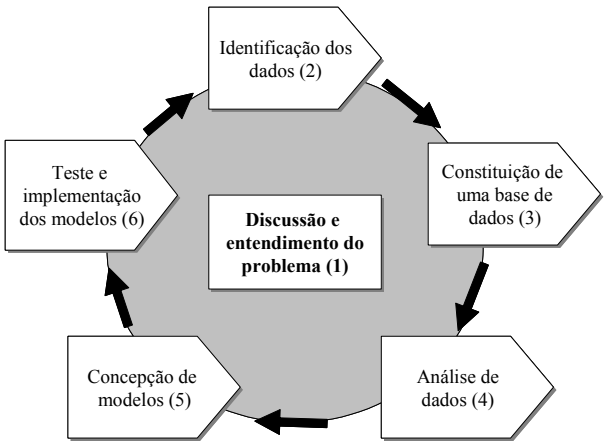
Ao se deter exclusivamente na dimensão modelos e técnicas de DM, a empresa corre o risco de se ver presa a eles. Certas empresas, por terem profissionais especializados na utilização de técnicas estatísticas e de análise de dados, e em razão de uma visão tecnicista da realidade, podem ser levadas a impor a ditadura da análise de dados. Nota-se que o sucesso na adoção de DM ocorre por meio de um processo evolutivo, pelo qual se caminha sempre nas três dimensões do processo. A partir do conhecimento do problema e da elaboração de uma estratégia de DM, caminha-se na utilização dos métodos de DM e ao mesmo tempo na exploração dos dados e no entendimento da sua composição e organização. A Figura 2 exibe o avanço progressivo nas três dimensões do processo.

5.1. Etapas do processo de DM

As etapas do processo de DM consistem em procedimentos que podem ser logicamente separados e organizados de maneira a facilitar a implementação e manutenção da aquisição de conhecimento.

Há seis etapas no processo de DM, como o indica a Figura 3:

Figura 3: As seis etapas do processo de DM



Fonte: Elaborada pelo autor.

Etapa 1: Discussão e entendimento do problema

Uma certa empresa do varejo tem uma base de dados de produtos e outra de clientes. O ponto de partida para um processo de DM bem-sucedido é a identificação adequada do problema, entendendo-se onde se quer chegar, que problema se quer resolver, ou que ação se quer criar. A empresa deverá identificar ações prioritárias e em que ordem ou de que maneira os produtos serão tratados e relacionados com o cliente, com o objetivo de oferecer os produtos e serviços certos para os clientes certos.

Etapa 2: Identificação dos dados

A identificação dos dados e a constituição de bases de dados permitirão que se inicie o processo de DM. A constituição da base de dados suscita algumas questões: (i) Quais variáveis são importantes para o problema em questão? (ii) Que quantidade de dados se deve coletar para fazer DM? (iii) Onde estão estes dados? (iv) Qual o custo para obtê-los? (v) Quem são os "donos" dos dados? De maneira geral, a resposta é: quanto mais dados, melhor. Em contrapartida, tem-se que considerar o custo de obtenção e análise desses dados.

Etapa 3: Constituição da base de dados

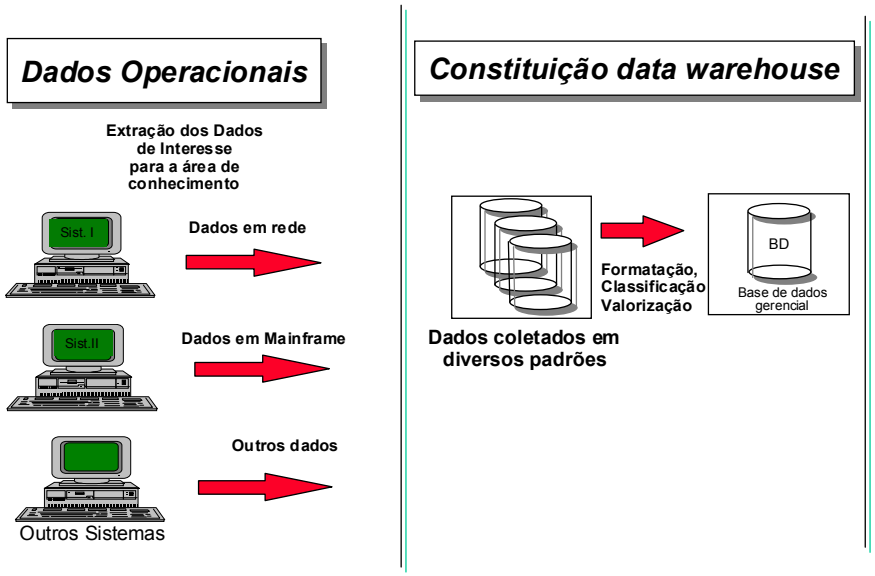
Uma origem importante dos dados para DM são, em geral, os sistemas que contêm os dados operacionais da empresa, que acumulam, por exemplo, o histórico do relacionamento com o cliente. Neste ponto destaca-se uma das etapas do ciclo do CRM, que é a organização dos dados em uma base de dados gerencial, em um *Data warehouse* (DW). O DW é uma coleção de dados orientada para assuntos, integrada, indexada pelo tempo e não-volátil, que apóia as decisões gerenciais, isto é, os dados contidos no DW são dados pré-processados a partir das bases de dados operacionais. Dessa forma, pode-se chamar os dados contidos em DW de dados gerenciais. O processo de criação de um DW é um dos primeiros pontos críticos do processo. Existem aspectos tecnológicos, organizacionais e comportamentais que dificultam o processo de implementação de uma base de dados gerencial em uma empresa (DE ALMEIDA, 1995; DE ALMEIDA, 1997; SILVA e FLEURY, 2000).

Tratando-se de uma grande empresa, os sistemas operacionais são normalmente fruto de vários esforços da empresa no seu processo de informatização. São freqüentes os casos nos quais sistemas operando em ambientes de informática distintos convivem. Na criação do DW, um esforço de conversão dos dados para um mesmo ambiente e num único padrão de dados é necessário. Isso faz com que a obtenção dos dados para a realização de um processo de DM seja um ponto crítico. O problema se torna mais complexo quando se está lidando com dados oriundos de diferentes áreas da organização. Departamentos que não interagem e áreas de informática com visões distintas são alguns dos problemas que podem ser encontrados pela organização na sua busca de dados para DM. A figura 4 ilustra o processo de criação de uma base de dados gerencial (*data warehouse*), envolvendo o tratamento de dados oriundos de diversos sistemas e em diversos padrões. O gerenciamento eficaz das relações entres as diversas áreas que deverão cooperar é determinante do sucesso do processo de organização dos dados (DE ALMEIDA, 1995; DE ALMEIDA, 1997; SILVA e FLEURY, 2000).

Podem ser ouvidas nas empresas frases como: "Nesta empresa a obtenção dos dados não será um problema. Eles estão todos em um mesmo sistema. É só extraí-los da base de dados". Descobre-se, entretanto, que os dados não estão organizados de maneira a atender às necessidades gerenciais, mas às demandas operacionais. Suponha-se o caso de uma empresa que quer usar as informações de movimentação de estoque para projetar o consumo de estoque como parte de um processo de geração de fluxo de caixa projetado. A empresa tem

informações de estoque armazenadas em seu sistema, mas, ao invés do histórico de consumo de um determinado item, o sistema armazenava somente o saldo atual do item. Para o controle de estoque, esta é uma informação essencial, pois permite saber se é o momento de fazer nova demanda do item. Mas para o processo de DM pode ser inútil, pois não informa sobre a dinâmica da utilização do item de estoque. Para projetar o consumo futuro, o histórico e a freqüência passada de consumo do estoque são informações indispensáveis.

Figura 4: O processo de criação de uma base de dados gerencial (*data warehouse*) envolve o tratamento de dados oriundos de diversos sistemas e em diversos padrões



Fonte: Elaborada pelo autor.

Etapa 4: Análise estatística de dados

O raciocínio estatístico permeia essa etapa e consiste em extração de informação generalizadora a partir de dados específicos, inferência, estimação, quantificação e redução da incerteza (DRANSFIELD, FISHER e VOGEL, 1999). Os fundamentos do raciocínio estatístico são a aleatoriedade, amostragem, análise de variabilidade e cálculo de probabilidades. Na prática, esse raciocínio se manifesta por meio de testes de hipóteses, especificação do modelo, estimação de parâmetros e construção de modelos. Um modelo estatístico, por sua vez, pode ser exploratório, confirmatório, preditivo ou explicativo. Os modelos são abstraídos a partir de dados. Os dados podem ser classificados como prontos para a análise (*ready data*), faltantes (*missing data*), discrepantes (*outlier*) ou desatualizados (*lagged data*). Portanto, como nem todos os dados estão prontos para a análise, torna-se necessário o seu pré-processamento, que consiste em selecionar as variáveis e a "janela de tempo" do estudo, planejar a amostragem, analisar os dados faltantes, defasados e discrepantes, imputar dados, converter dados defasados e/ou discrepantes em faltantes, transformar os dados (normalização/simetrização e padronização), recodificar os dados, atribuir identificadores adequados, reduzir a dimensionalidade dos dados por meio de análise de correlações e de análise de componentes principais, etc. Além disso, algumas estatísticas são utilizadas para medir a similaridade entre os clientes, tais como: qui-quadrado, correlação, estatística t, estatística F, entropia e a divergência de Kullback-Leibler. O pré-processamento consome em geral mais da metade do tempo de DM. No campo das ferramentas e técnicas para DM, pode-se classificar as abordagens de DM em dois tipos, segundo a maneira como os dados serão explorados: (1)

abordagens baseadas nos conceitos e hipóteses da estatística e (2) abordagem baseada nos dados. No primeiro caso, trata-se do uso de técnicas estatísticas convencionais, que supõem premissas ou hipóteses iniciais sobre os dados; a partir de então, as técnicas são usadas. Utilizam-se modelos cuja escolha significa assumir um certo comportamento em relação aos dados que permita usá-los (o que é chamado de hipóteses sobre a distribuição dos dados para que as técnicas possam funcionar adequadamente). O cálculo da média dos indivíduos, por exemplo, tem pouco significado se a distribuição dos dados (que mostra a freqüência, ou o número de indivíduos em diferentes faixas de indivíduos para uma determinada variável) não for simétrica (HAIR JR. *et al.*, 1998).

Uma segunda maneira de caminhar é não assumir padrão algum de distribuição nos dados e usar métodos que criam modelos explicativos ou preditivos a partir da própria interação com os dados. São os métodos baseados na "força bruta" computacional, pois exigem forte iteração computacional: árvore de classificação e regressão, redes neurais e reamostragem são os exemplos mais importantes nessa categoria. Ao contrário da abordagem anterior, estas exigem poucas hipóteses prévias para que se possa utilizá-las. As duas abordagens não são excludentes, mas, ao contrário, complementares, e na fase de análise dos dados ambas são empregadas.

Na etapa 4 começa o processo de exploração dos dados em busca de conhecimento que possibilite o contato com o problema através dos dados. Os dados coletados são explorados com o objetivo de avaliar seu potencial como fonte bruta de conhecimento sobre o problema. Nesta fase é explorada a capacidade dos dados de identificar grupos (análise de grupos), das variáveis em classificar corretamente indivíduos nos grupos respectivos (análise discriminatória e classificatória), de sua capacidade de previsão, de identificar relações entre as variáveis, entre os produtos, etc. Pode-se dizer que a etapa 4 é a etapa de "separar o joio do trigo", isto é, separar variáveis que interessam das que não têm poder preditivo ou discriminatório, ou seja, não têm informação útil a respeito do problema. Este é o momento no qual se descobre a qualidade dos dados de que se dispõe, do ponto de vista de DM. É uma fase que também pode ser chamada fase de pré-processamento dos dados, pois é quando se descobre o valor informacional da base de dados (*e.g.*, será que patrimônio é uma variável interessante para separar os clientes com propensão à compra daqueles não propensos à compra de determinado produto?). Descobrem-se variáveis úteis e variáveis com baixo poder explicativo ou preditivo para o problema. Segundo HAIR Jr. *et al.* (1998), há um aspecto no qual as técnicas de DM se assemelham, sejam elas uma análise de regressão ou uma rede neural: se entrar "lixo", sairá lixo! Esta é a fase de tirar o "lixo" dos dados, para que no final se possa chegar a conclusões úteis. Nem a técnica mais poderosa resiste à entrada de lixo. Se for criado um modelo de redes neurais a partir de dados quaisquer, obtém-se um modelo que dá respostas quaisquer. Ainda, segundo os autores, mesmo as redes neurais não têm a capacidade de transformar dados de baixa qualidade ou com distorções significativas em modelos satisfatórios. Os dados devem ser examinados cuidadosamente, levando-se em consideração a amostragem adequada, a distribuição dos dados e sua transformação. Para que se possa obter o máximo de informação e conhecimento dos dados, é importante fazer análises descritivas cuidadosas, procurando-se conhecer o padrão subjacente a eles, isto é, é preciso ganhar "intimidade" com os dados.

As etapas 5 e 6 do processo de DM são dedicadas aos modelos. Neste momento, são utilizadas as técnicas de modelagem do conhecimento: redes neurais, árvore de classificação e regressão, análise de regressão, etc.

Etapas 5 e 6: Concepção, teste e implementação dos modelos de DM

O processo de DM busca primordialmente os seguintes resultados: descrição e visualização, classificação, estimação, previsão, agrupamento e regras de associação (BERRY e LINOFF, 2000). Classificação, estimação e previsão dizem respeito à criação de modelos que servirão para identificar o comportamento de novos clientes. Será que determinado cliente irá comprar tal produto em resposta a uma mala direta? (previsão). Determinado cliente é um "bom" ou "mal" cliente? (classificação). Qual o risco de inadimplência de determinado cliente para uma concessão de crédito? (estimação).

Descobre-se através de uma análise de grupos que um determinado agrupamento de clientes tem características comuns. Com a análise de grupos

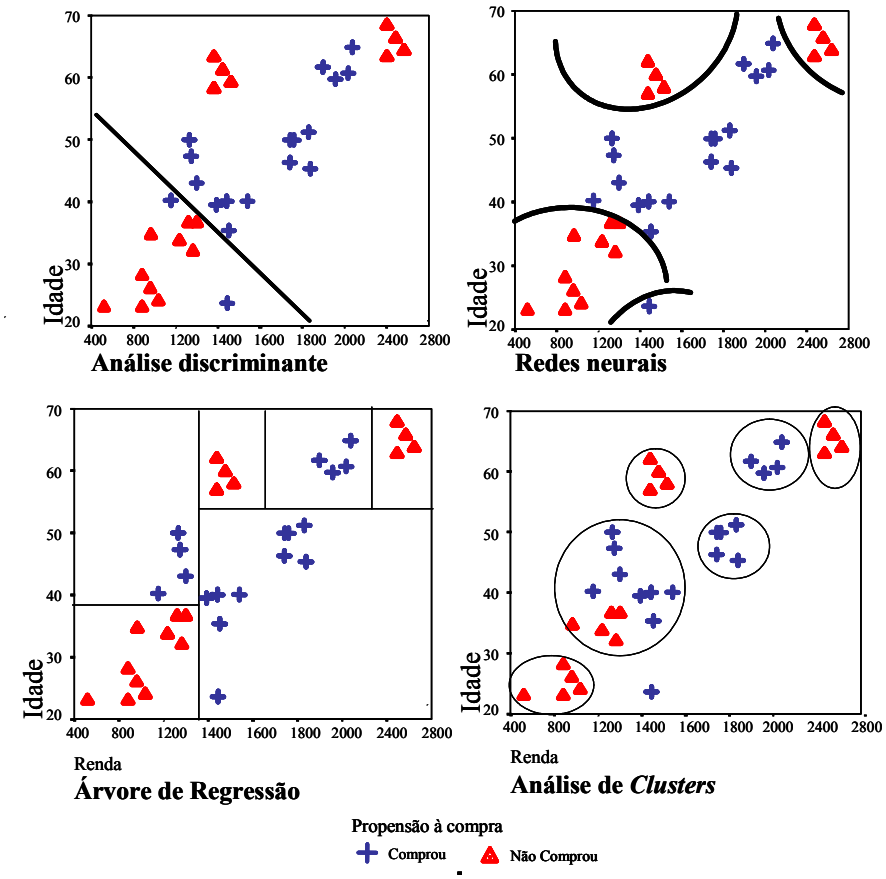
identificam-se grupos que se diferenciam signficativamente dos clientes dos outros grupos (e.g. clientes de alto poder aquisitivo, que gostam de esporte, viajam muito e têm menos de 40 anos). Ou descobre-se que quem compra o produto A também se interessa em comprar o produto B (regra de associação).

6. TECNOLOGIA PARA DM

O processo de DM procura associar a existência de dados abundantes à utilização de tecnologia de informática para formalizar conhecimento (DE ALMEIDA, 1995; BERRY e LINOFF, 1997; HAIR JR. *et al.*, 1998). Os dados acumulados por uma organização podem ser encarados como um conhecimento tácito, que pode ser explicitado e estruturado a partir do uso de novas tecnologias de informática disponíveis atualmente, que operacionalizam as técnicas que comentamos anteriormente. Diversos são os modelos e técnicas que podem auxiliar a organização a transformar conhecimento tácito em conhecimento formalizado e disponível, utilizando-se intensamente de recursos e da potência de processamento da informática (BERRY e LINOFF, 1997; HAIR JR. *et al.*, 1998).

FLORES (1998) classifica as novas técnicas de formalização do conhecimento, que servem para compreender e antecipar, como máquinas estatísticas (algoritmos) que aprendem. Essas técnicas procuram, por meio de diferentes algoritmos, criar estruturas de conhecimento a partir de dados disponíveis. Uma grande quantidade de técnicas desse tipo tem surgido (ROSENFELD e WECHSLER, 2000). FLORES (1998) identifica quatro técnicas básicas (ilustradas na figura 5) para a criação de modelos preditivos de DM: análise discriminatória e classificatória, árvores de regressão e classificação, redes neurais e análise de grupos.

Figura 5: Quatro técnicas básicas de DM



Fonte: Adaptado de FLORES (1998).

Como exemplo, pretende-se separar as pessoas físicas de uma base de dados a partir de duas características básicas, idade (ano) e renda (R$), e avaliar a sua propensão à compra de um seguro residencial. As técnicas de análise de grupos, *e.g.*, redes neurais artificiais de Kohonen, separam as pessoas físicas em grupos usando como dados algumas características comuns relevantes. Para a criação dos grupos são calculadas as distâncias entre as pessoas físicas. Essas distâncias podem ser formalizadas de diferentes maneiras (HAIR JR. *et al.*, 1998). Conforme a figura 5, pode-se observar que há um grupo de pessoas físicas ao alto e à direita que não é propenso à compra do seguro: são os indivíduos de mais avançada idade e de maior renda. A questão importante aqui é entender o motivo pelo qual esse grupo se diferencia dos outros. Talvez sejam pessoas aposentadas, que preferem se ocupar por conta própria da segurança da casa. Mas não está disponível a informação sobre a situação de trabalho das pessoas físicas dessa coleção de dados.

Outra maneira de tratar esses dados é usar uma técnica de DM para classificar as pessoas físicas em grupos que já são conhecidos de antemão: propensos ou não à compra do seguro. As técnicas discriminatórias e classificatórias normalmente fazem separações lineares ou quadráticas das pessoas físicas em grupos. As redes neurais separam as pessoas físicas em grupos de maneira não-linear (*vide* WONG, BODNOVICH e SELVI, 1997).

A árvore de regressão e classificação pode ser representada graficamente por um espaço cartesiano das variáveis divididas por retângulos (*vide* DUDA *et al.*, 2001). Esta técnica segue basicamente o seguinte princípio: é selecionada dentre as variáveis aquela que tem maior poder de separação dos dois grupos. Uma vez feita a separação das pessoas físicas, uma segunda variável é escolhida e os subgrupos são novamente divididos em células retangulares menores. E assim sucessivamente, até que se atinja uma divisão satisfatória. Uma variável pode ser utilizada mais de uma vez para criar uma nova subdivisão. É o caso do exemplo da figura 5, na qual foram geradas 6 células retangulares a partir de duas variáveis: três grupos de compradores (sinal de adição) e três de não compradores (triângulo).

As técnicas estatísticas discriminatórias e classificatórias lineares se comportam bem quando as variáveis possuem uma distribuição elíptica multivariada homocedástica, grande quantidade de pessoas físicas e matriz de dados completa. Já as árvores de regressão e classificação são interessantes no caso em que existem poucos grupos a serem identificados (*e.g.*, comprou ou não comprou), pequena quantidade de pessoas físicas, grande quantidade de variáveis e matriz de dados incompleta. Modelos de regressão podem ser utilizados em combinação com as árvores de regressão e classificação para aumentar a capacidade preditiva (*vide* SHEPARD, 1999, e ANSWERTREE, 2001). Redes neurais são mais eficazes que uma árvore de regressão e classificação quando se trata de estimar variáveis dependentes quantitativas/métricas tais como valor de vendas, demandas, fluxo de caixa, preços, etc. (ADYA e COLLOPY, 1998; CORRÊA e PORTUGAL, 1998; SILVA, PORTUGAL e CECHIN, 2001; DIAZ e ARAÚJO, 1998).

Para uma comparação entre essas técnicas, *vide* FLORES (1998), BERRY e LINOFF (1997, 2000), BERSON *et al.* (2000) e HAIR JR. *et al.* (1998). Existem alguns *softwares* de DM no mercado (WESTPHAL e BLAXTON, 2000). Alguns deles são bastante específicos e exploram intensamente uma técnica em particular, tais como *softwares* que exploram especificamente redes neurais (*e.g.*, SNNS). Existem *softwares* projetados especificamente para DM, tais como Entreprise Miner (SAS), Intelligent Miner (IBM) e Clementine (SPSS).

7. CONCLUSÕES

Data mining foi abordada neste texto, tanto do ponto de vista do processo e de seus aspectos de implementação, quanto das técnicas, de suas características e de sua utilização. Sua utilização envolve uma complexidade tanto técnica quanto organizacional, ou mesmo operacional. Foi discutida a existência de três dimensões que envolvem a dinâmica do processo de DM, produtos e serviços, modelos e técnicas, dados e bases de dados. Percebeu-se, na empresa estudada, que as três dimensões se desenvolvem de maneira evolutiva e interativa e que não é interessante a tomada de decisão sobre uma das três dimensões sem a reflexão sobre as outras duas.

Essas dimensões interagem com 6 etapas do processo de DM que foram destacadas neste texto,

mostrando outros aspectos do processo. Foram destacadas neste texto as técnicas estatísticas utilizadas em DM, pois se trata de um ponto complexo e importante no processo.

As reflexões aqui apresentadas são fruto de trabalhos realizados principalmente numa empresa, porém com contribuições de outras intervenções realizadas pelos autores. No entanto, o modelo de processo proposto pode ser mais aprofundado à luz de outras experiências.

Um ponto pouco explorado neste texto, que merece atenção especial, é o aspecto humano dos processos de DM, em particular a discussão sobre o perfil do novo profissional de DM, que é discutido em outros trabalhos (*vide* DRANSFIELD, FISHER e VOGEL, 1999; HAHN e HOERL, 1998).

8. REFERÊNCIAS BIBLIOGRÁFICAS

ADYA, M.; COLLOPY, F. How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, 17, p. 481-495, 1998.

ANSWERTREE 3.0: *User's Guide*. SPSS Inc., 2001.

BERRY, M. J. A.; LINOFF, G. *Data mining Techniques for Marketing*, *Sales and Customer Support*. N.Y.: Wiley, 1997.

BERRY, M. J. A.; LINOFF, G. *Mastering Data mining*. N.Y.: Wiley, 2000.

BERSON, A.; SMITH, S.; THEARLING, K. *Building Data mining Applications for CRM*. McGraw-Hill, 2000.

BRETZKE, M. *Marketing de relacionamento e competição em tempo*. São Paulo: Atlas, 2000.

CORRÊA, W. R.; PORTUGAL, M. S. Previsão de séries de tempo na presença de mudança estrutural: redes neurais artificiais e modelos estruturais. *Revista de Economia Aplicada*, v. 2, n. 3, p. 487-513, 1998.

DE ALMEIDA, F. C. Modeling and Implementing a Long Term Cash Flow System: An Executive Information Systems Approach. In: CONFERENCE ON DECISION SUPPORT SYSTEMS, 4th, 1997, Lausanne. *Anais*... Lausanne, 1997. p 489-502.

DE ALMEIDA, F. C. Atores e Fatores na Introdução de um Sistema de Informação. *Revista Brasileira de Administração Contemporânea*, v. 1, n. 4, p. 177-192, 1995.

DIAZ, M. D. M.; ARAÚJO, L. J. S. Aplicação de redes neurais à economia: demanda por moeda no Brasil. *Revista de Economia Aplicada*, v. 2, n. 2, p. 271-297, 1998.

DRANSFIELD, S. B.; FISHER, N. I.; VOGEL, N. J. Using Statistics and statistical thinking to improve organisational performance. *International Statistical Review*, v. 67, n. 2, p. 99-150, 1999.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. 2. ed. N.Y.: Wiley, 2001.

FLORES, J. G. *Statistical and Machine Learning Frameworks for Economics*: Analysis of Error Curves and Applications to Derivatives Pricing and Credit Risk Assessment. Tese (Doutorado). Harvard University, 1998.

HAHN, G.; HOERL R. Key challenges for Statisticians in Business and Industry. *Technometrics*, v. 40, n. 3, p. 195-213, 1998.

HAIR JR., J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. *Multivariate Data Analysis*. 5th ed. N.J.: Prentice Hall, 1998.

PEPPERS, W; ROGERS, M.; DORF, B. *One to One Field Book*. N.Y: Currency Book, 1999.

REICHHELD, R. *A Estratégia da Lealdade*. R.J.: Campus, 1996.

ROSENFELD, A.; WECHSLER, H. Pattern recognition: historical perspective and future directions. *International Journal of Imaging Systems and Technology*, n. 11, p. 101-116, 2000.

SHEPARD, David & Associates. *The New Direct Marketing:* How to Implement a Profit-Driven Database Marketing Strategy. 3. ed. McGraw-Hill, 1999.

SCHÜRMANN, J. *Pattern classification*: a unified view of statistical and neural approaches. N.Y.: Wiley, 1996.

SILVA, A. B. M.; PORTUGAL, M. S.; CECHIN, A. L. Redes Neurais artificiais e análise de sensibilidade: uma aplicação à demanda de importações brasileira. *Revista de Economia Aplicada*, v. 5, n. 4, p. 645-693, 2001.

SILVA, S. M; FLEURY, M. T. Aspectos culturais do uso de tecnologias de informação em pesquisa acadêmica. *Revista de Administração da USP* (RAUSP), São Paulo: Universidade de São Paulo, v. 35, n. 2, 2000.

VAVRA, T. G. *After-Marketing*. Irwin, 1995.

WEIR, J. A Web/Business Intelligence solution. *Information Systems Management*, v. 7, n. 1, p. 441-46, 2000.

WESTPHAL, C.; BLAXTON, T. *Data mining Solutions*. N.Y.: Wiley, 2000.

WITTEN, I. H.; FRANK E. *Data mining*: Practical Machine Learning Tools and Techniques with JAVA Implementations. S.F.: Morgan Kaufmann, 2000.

WONG, B. K.; T. A. BODNOVICH, T. A.; SELVI, Y. Neural network applications in business: A review and analysis of the literature (1988-1995). *Decision Support Systems*, n. 19, p. 301-320, 1997.

9. OBRAS CONSULTADAS

ANSWERTREE. *Treinamento*. SPSS Brasil Ltda, 2000.

BOCK, H.-H.; DIDAY, E. *Analysis of Symbolic Data*: Exploratory Methods for Extrating Statistical Information from Complex Data. Berlim: Springer, 2000.

DE ALMEIDA, F. C. *L'Evaluation des risques de défaillance des entreprises à partir des réseaux de neurones insérés dans les systèmes d'aide à la décision*. Tese (Doutorado em Administração). École Supérieure des Affaires Grenoble: Universidade de Grenoble, 1993.

GARVER, M. S. Using data mining for customer satisfaction research. *Marketing Research*, Chicago, v. 14, n. 1, p. 8-12, 2002.

JOHNSON, R.; WICHERN, D. *Applied Multivariate Statistical Analysis*. 5th ed. N.J.: Prentice Hall, 2002.

KOHONEN, T. *An Introduction to Neural Computing*. Neural Networks, n. 1, p. 16, 1988.

KOHONEN, T. *Self-organization and Associative Memory*. Berlim: Springer-Verlag, 1984.

NEURAL CONNECTION. *Applications Guide*. SPSS Inc., 1997.

PRAHALAD, C. K.; HAMEL, G. "The Core Competence of the Corporation". *Harvard Business Review*, v. 68, n. 3, p. 79-91, May- June 1990.

PRINCIPE, J. C.; EULIANO, N. R.; LEFEBVRE, W. C. *Neural and Adaptive Systems*. N.Y.: Wiley, 2000.

RUMELHART, D. E.; McCLELLAND, J. C.; PDP Research Group. *Parallel Distributed Processing - Exploration in the Microtexture of Cognition*. Londres: MIT, 1986. v. 1.