Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining

Eliane Prezepiorski Lemos
Maria Teresinha Arns Steiner
Julio César Nievola

RESUMO

Na área de crédito bancário, a posse e o uso de ferramentas que auxiliem na tarefa de classificação de clientes em prováveis solventes ou insolventes em relação à tomada de crédito podem tornar-se um fator-chave, resultando em uma grande vantagem competitiva. Dentro da imensa quantidade de dados disponíveis nos bancos de dados das empresas existe muito conhecimento útil e importante que está **escondido**. Com a metodologia de *Data Mining*, pode-se transformar esses dados em informações valiosas para auxiliar no processo decisório. Neste trabalho são analisados registros históricos de 339 clientes (pessoas jurídicas) de uma agência bancária, por meio de duas das ferramentas de *Data Mining*: Redes Neurais e Árvores de Decisão. Essas técnicas permitem fazer o reconhecimento de padrões e também classificar novos casos. Os resultados foram bastante satisfatórios, mostrando que, para esse problema específico, as Redes Neurais apresentaram uma taxa de classificação correta maior do que aquela das Árvores de Decisão.

Palavras-chave: KDD, *data mining*, árvores de decisão, redes neurais, crédito bancário.

Recebido em 31/março/2004
Aprovado em 01/abril/2005

1. INTRODUÇÃO

O ambiente de negócios atual está muito competitivo. Especificamente no caso de crédito bancário, a posse e o uso de ferramentas que auxiliem na tarefa de classificação de clientes em prováveis solventes e insolventes são um fator-chave, que resulta numa grande vantagem competitiva.

A partir da década de 1980, passou-se a ter maior facilidade de acesso a equipamentos computacionais e houve uma queda no custo de armazenamento de dados. Com isso, algumas empresas verificaram que, em meio a essa imensa quantidade de dados, informações valiosas poderiam estar escondidas. Em função dessa percepção, elas passaram a buscar meios de transformar os da-

Eliane Prezepiorski Lemos, Mestre em Métodos Numéricos em Engenharia na área de Programação Matemática pela Universidade Federal do Paraná, é Professora do Departamento de Matemática da UNICENTRO (CEP 85010-990 — Guarapuava/PR, Brasil) e funcionária do Banco do Brasil S.A., Agência Guarapuava, Paraná.
E-mail: elemos@unicentro.br
Endereço:
UNICENTRO
Departamento de Matemática
Rua Presidente Zacarias, 875
85010-990 — Guarapuava — PR

Maria Teresinha Arns Steiner, Mestre e Doutora em Engenharia de Produção na área de Pesquisa Operacional pela Universidade Federal de Santa Catarina, é Professora Adjunto IV do Departamento de Matemática e do Programa de Pós-Graduação em Métodos Numéricos em Engenharia da Universidade Federal do Paraná (CEP 81531-990 — Curitiba/PR, Brasil).
E-mail: tere@mat.ufpr.br

Julio César Nievola, Mestre e Doutor em Engenharia Elétrica na área de Sistemas de Informação pela Universidade Federal de Santa Catarina, é Professor Titular do Curso de Ciência da Computação e do Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná (CEP 80215-901 — Curitiba/PR, Brasil).
E-mail: nievola@ppgia.pucpr.br

dos em informação. Isso acontece quando o dado ganha um significado para seu utilizador; caso contrário, continua sendo simplesmente um dado.

A Mineração de Dados (*Data Mining*) é uma nova metodologia para melhorar a qualidade e a eficiência das decisões, por meio da obtenção de conhecimento útil para tomada de decisões estratégicas, baseada nos dados históricos armazenados. Segundo Steiner *et al*. (1999, p.56), "a correta decisão de crédito é essencial para a sobrevivência das empresas bancárias". Afirmam ainda que "qualquer erro na decisão de concessão pode significar que em uma única operação haja a perda do ganho obtido em dezenas de outras bem-sucedidas". O que é desejável e necessário, então, é "analisar uma proposta de negócio e comparar o **custo de conceder** com o **custo de negar** a operação" (STEINER *et al.*, 1999, p.56).

Ao fazer-se o correto uso de ferramentas na análise de crédito, várias são as vantagens obtidas, dentre as quais podem ser destacadas:

- necessidade de menos pessoas envolvidas com a análise do crédito, as quais podem ser aproveitadas em outras atividades;
- maior rapidez no processamento dos pedidos de crédito;
- menor subjetividade no processo;
- direcionamento mais eficaz do crédito.

O objetivo principal neste trabalho é analisar o uso de duas metodologias para a classificação de empresas como sendo adimplentes ou inadimplentes. Dentre as diversas técnicas de Mineração de Dados existentes — Análise de *Cluster*, Árvores de Decisão, Redes Neurais, Indução de Regras, Algoritmos Genéticos, Aprendizado Baseado em Casos —, optou-se pela utilização de Redes Neurais e Árvores de Decisão.

A intenção de utilizar duas técnicas é fazer a comparação dos resultados obtidos em cada uma delas, procurando determinar qual oferece a maior taxa de acerto para o contexto do presente trabalho, ou seja, na classificação de novas empresas como prováveis adimplentes ou inadimplentes.

2. DESCRIÇÃO DO PROBLEMA REAL

Os dados utilizados neste artigo são reais e foram obtidos no Banco do Brasil S.A., Agência Guarapuava (Paraná), que detém uma grande fatia do mercado de pessoas jurídicas da cidade, no que diz respeito ao crédito bancário. O Banco do Brasil procura atender às necessidades desse seg-

Informações Referentes à Amostra

Código	Informações	Valor
A	Existência de restrições em nome da empresa	1 = Sim 2 = Não
B	Existência de restrições baixadas nos últimos cinco anos em nome da empresa	1 = Sim 2 = Não
C	Tempo de conta no Banco do Brasil	Valor numérico em meses
D	Setor de atividade	1 = Comércio 2 = Indústria 3 = Serviços
E	Tempo de atividade	1 = Mais de 9 anos 2 = De 6 a 9 anos 3 = De 3 a 5 anos 4 = De 1 a 2 anos 5 = Menos de 1 ano
F	Número de funcionários	Valor numérico
G	Sede da empresa (imóvel)	1 = Próprio 2 = Alugado 3 = Cedido
H	Nome do bairro	1 = Centro 2 = Outros
I	Principais clientes	1 = Pessoas físicas 2 = Pessoas jurídicas 3 = Misto
J	Faturamento bruto anual	Valor numérico
K	Cliente em outro banco	1 = Sim 2 = Não
L	Bens imóveis	Valor numérico
M	Bens móveis	Valor numérico
N	Seguro empresarial	1 = Sim 2 = Não
O	Aplicações financeiras no Banco do Brasil	1 = Sim > 8.000 2 = Sim 4.000 a 8.000 3 = Sim 2.000 a 4.000 4 = Sim < 2.000 5 = Não
P	Vendas a prazo	1 = Menos de 20% 2 = Mais de 20%
Q	Experiência de crédito no Banco do Brasil	1 = Sim > 2 anos 2 = Sim < 2 anos 3 = Não
R	Histórico da conta corrente	1 = Normal 2 = Cheques devolvidos 3 = Cliente novo 4 = Pequenos atrasos freqüentes
S	Sócios da empresa possuem restrições	1 = Sim 2 = Não
T	Sócios da empresa tiveram restrições baixadas nos últimos cinco anos	1 = Sim 2 = Não
U	Sociedade entre cônjuges	1 = Sim 2 = Não
V	Existência de bens imóveis em nome dos sócios	Valor numérico
W	Existência de bens móveis em nome dos sócios	Valor numérico
X	Risco atribuído pelo Banco do Brasil*	1 = A 2 = B 3 = C 4 = D 5 = E
Y	Resultado	1 = Adimplente 2 = Inadimplente

* Essa variável é um conceito definido pelo aplicativo ANC, por meio do qual são estipuladas as garantias mínimas exigidas nas operações de crédito. Na escala, A é o melhor e E é o pior conceito.

mento de mercado, colocando à disposição linhas de crédito tanto para capital de giro quanto para investimentos. Além disso, a clientela da carteira de pessoa jurídica do Banco do Brasil S.A. é constituída tanto de micro e pequenos como também de médios empresários. As grandes empresas não fazem parte da carteira da agência em questão, tendo em vista que o Banco possui agências chamadas Empresariais, especializadas para esse público.

Atualmente, o Banco do Brasil utiliza, como ferramenta para realizar sua análise de crédito, um aplicativo interno chamado Análise de Crédito (ANC). É desse aplicativo, que contém as informações cadastrais e contábeis das empresas, que a gerência do Banco se vale para apoiar suas decisões de conceder ou não crédito bancário.

Neste trabalho, utilizaram-se os dados históricos de 339 clientes pessoa jurídica da referida Agência, dos quais 266 são adimplentes e 73 inadimplentes. De cada um deles foram extraídas as 24 informações especificadas no quadro da página 226, para as quais os valores são os indicados.

3. DESCOBERTA DE CONHECIMENTO E *DATA MINING*

As técnicas e ferramentas que buscam transformar os dados armazenados nas empresas em conhecimento são o objetivo da área denominada Descoberta de Conhecimentos em Bases de Dados (*Knowledge Discovery in Databases* — KDD).

Segundo Fayyad *et al.* (1996), o termo KDD foi criado em 1989 como referência ao processo amplo de encontrar conhecimento em dados. KDD refere-se a todo processo de descoberta de conhecimento útil de dados, enquanto Mineração de Dados refere-se à aplicação de algoritmos para extrair modelos dos dados. Até 1995, muitos autores consideravam os termos KDD e Mineração de Dados como sinônimos.

O processo de KDD é um conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados. Segundo Fayyad *et al.* (1996), esse conjunto é composto por cinco etapas: seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; Mineração de Dados; interpretação e avaliação dos resultados. Essas etapas podem ser visualizadas na figura 1.

O processo de KDD começa, obviamente, com o entendimento do domínio da aplicação e dos objetivos finais a serem atingidos. Em seguida, é feito um agrupamento organizado da massa de dados que é alvo da prospecção. A etapa da limpeza dos dados (*Data Cleaning*) vem a seguir, por um pré-processamento dos dados, visando adequá-los aos algoritmos. Isso se faz por meio de integração de dados heterogêneos, eliminação de incompletude dos dados e outras operações. Essa etapa pode tomar até 80% do tempo necessário de todo o processo, devido às dificuldades de integração de bases de dados heterogêneas (MANNILA, 1996).

Os dados pré-processados devem ainda passar por uma transformação para o seu armazenamento adequado, visando facilitar o uso das técnicas de Mineração de Dados. Atualmente, o uso de depósitos ou armazéns de dados (*Data Warehouse*) está se tornando bastante intenso, já que com essa tecnologia as informações são armazenadas de maneira mais eficiente. Segundo Inmon (1997), o armazém de dados é um conjunto de dados, integrado, não volátil nem variável em relação ao tempo, que dá apoio às decisões gerenciais.

Prosseguindo no processo, chega-se à fase de Mineração de Dados especificamente, que começa com a escolha das ferramentas a serem utilizadas. Essa escolha depende fundamental-

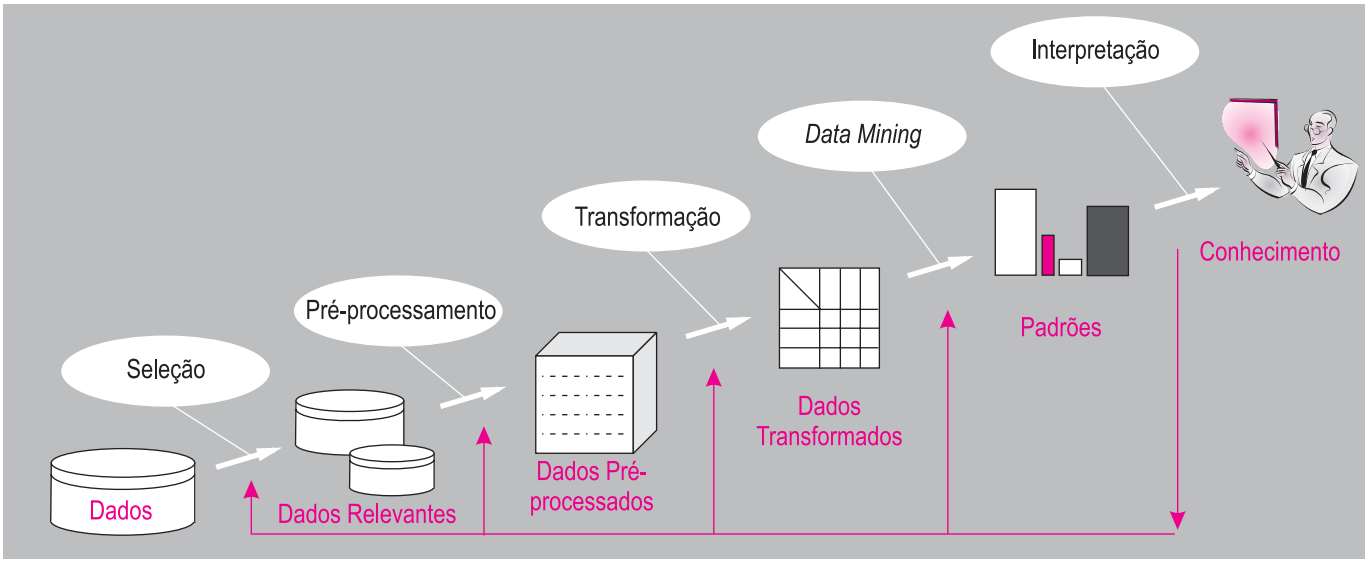

Figura 1: Processo de KDD

Fonte: Fayyad *et al.* (1996).

mente do objetivo do processo de KDD (WITTEN e FRANK, 2000): classificação, agrupamento ou associação. De modo geral, na fase de Mineração de Dados, técnicas especializadas procuram padrões nos dados.

Diversas técnicas distintas, como redes neurais, árvores de decisão, sistemas baseados em regras e programas estatísticos, tanto isoladamente quanto em combinação, podem ser aplicadas ao problema. Em geral, o processamento de busca é interativo, de forma que os analistas revêem o resultado, formam um novo conjunto de questões para refinar a busca em um determinado aspecto da descoberta e realimentam o sistema com novos parâmetros. Ao final do processo, o sistema de Mineração de Dados gera um relatório das descobertas, que passa então a ser interpretado pelos analistas de Mineração. Somente após a interpretação das informações obtidas é que se encontra conhecimento. Mineração de Dados é a parte mais interessante do processo de KDD e, no contexto de negócios, a fase que mais alavanca e auxilia o empresário a descobrir filões de mercado.

O cérebro humano consegue trabalhar com algo em torno de sete unidades de informação ao mesmo tempo. A função da Mineração de Dados é justamente ampliar essa quantidade de informações simultâneas e tornar isso visível ao olho humano (POSSAS *et al*., 1998). "Mineração de Dados é uma ferramenta utilizada para descobrir novas correlações, padrões e tendências entre as informações de uma empresa, através da análise de grandes quantidades de dados armazenados em armazéns de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas" (NIMER e SPANDRI, 1998, p.32).

Técnicas de Mineração de Dados têm sido aplicadas com sucesso para a solução de problemas em diversas áreas, como:

- **Vendas** — buscando a retenção de clientes, ou seja, identificando clientes que podem migrar para o concorrente e tentar retê-los; detectar associações entre produtos; identificar padrões de comportamento de consumidores; encontrar características dos consumidores de acordo com a região demográfica; prever quais consumidores serão atingidos nas campanhas de *marketing* e, nesses casos, enviar mala-direta anunciando o produto apenas para aqueles prováveis compradores, tarefa denominada de mala-direta direcionada.

- **Finanças** — detectar padrões de fraudes no uso dos cartões de crédito; identificar os consumidores que estão tendendo a mudar de companhia de cartão de crédito; identificar regras a partir dos dados do mercado; encontrar correlações escondidas nas bases de dados.

- **Seguros e planos de saúde** — determinar os procedimentos médicos requisitados ao mesmo tempo; prever quais consumidores têm tendência a comprar novas apólices; identificar comportamentos fraudulentos.

- **Transporte** — determinar a distribuição dos horários entre nos vários caminhos; analisar padrões de sobrecarga.

- **Medicina** — caracterizar o comportamento dos pacientes para prever novas consultas; identificar terapias de sucesso para diferentes doenças; prever quais pacientes têm maior probabilidade de contrair uma certa doença, em função de dados históricos de pacientes e doenças.

- **Telecomunicação** — identificar fraudes em ligações telefônicas dentre um número de ligações efetuadas pelos clientes.

- **Mercado financeiro** — prever as ações que irão subir ou descer na bolsa de valores, em função de dados históricos com preços de ações e valores de índices financeiros.

Segundo Freitas (2000), o conhecimento a ser descoberto deve satisfazer a três propriedades: ser correto (tanto quanto possível); ser compreensível pela maioria dos usuários; e ser interessante, útil, novo e surpreendente. Ainda segundo Freitas (2000), o método de descoberta do conhecimento deve apresentar as seguintes características: ser eficiente, genérico (ou seja, aplicável a vários tipos de dados) e flexível (facilmente modificável).

Dentre as técnicas de Mineração de Dados utilizadas para classificação, destacam-se as Árvores de Decisão e as Redes Neurais, que são abordadas e aplicadas ao problema descrito na seção 2 deste artigo.

A Árvore de Decisão é um método adequado quando o objetivo da Mineração de Dados é a classificação de dados ou predição de saídas. É conveniente usar Árvore de Decisão quando o objetivo for a categorização dos dados. Ela também é uma boa escolha quando o objetivo é gerar regras que podem ser facilmente entendidas, explicadas e traduzidas para linguagem natural.

As Redes Neurais tentam construir representações internas de modelos ou padrões detectados nos dados, as quais geralmente não são visíveis ao usuário. As Redes Neurais utilizam um conjunto de elementos de processamento (nós) análogos aos neurônios. Esses elementos de processamento são interconectados em uma rede que pode identificar padrões nos dados, ou seja, a rede aprende através da experiência, tal como as pessoas (DIN, 1998).

O uso de Mineração de Dados para a construção de um modelo deve trazer as seguintes vantagens:

- os modelos devem ser de fácil compreensão, ou seja, pessoas sem conhecimento estatístico devem poder interpretar o modelo e compará-lo com as próprias idéias, permitindo ao usuário obter mais conhecimento sobre o comportamento do cliente e a possibilidade de usar essa informação para otimizar os processos dos negócios;
- permitir que grandes bases de dados possam ser analisadas, ou seja, grandes conjuntos de dados, de até vários *gigabytes*

de informação, podem ser analisados com Mineração de Dados para descobrir informações até então desconhecidas;
- as variáveis não devem necessitar de recodificação, já que se deve permitir trabalhar tanto com variáveis numéricas (quantitativas) quanto com variáveis categóricas (qualitativas), devendo elas aparecer no modelo exatamente da mesma forma em que aparecem na base de dados;
- os modelos devem ser precisos, permitindo a validação por técnicas estatísticas.

4. ÁRVORES DE DECISÃO E REDES NEURAIS

Nesta seção é apresentada uma descrição mais detalhada das técnicas abordadas.

4.1. Árvores de decisão

Árvores de Decisão são métodos de classificação de dados no contexto da chamada Mineração de Dados (*Data Mining*). Podem ser usadas em conjunto com a tecnologia de indução de regras, mas são as únicas a apresentar os resultados hierarquicamente (com priorização). Nelas, o atributo mais importante é apresentado na árvore como o primeiro nó, e os atributos menos relevantes são mostradas nos nós subseqüentes. A vantagem principal das Árvores de Decisão é a **tomada de decisões** levando em consideração os atributos mais relevantes, além de compreensíveis para a maioria das pessoas. Ao escolher e apresentar os atributos em ordem de importância, as Árvores de Decisão permitem aos usuários conhecer quais fatores mais influenciam os seus trabalhos.

Quinlan (1993) desenvolveu uma técnica que permitiu o uso da representação do conhecimento por meio das Árvores de Decisão. Sua contribuição consistiu na elaboração de um algoritmo chamado ID3 que, juntamente com suas evoluções (ID4, ID6, C 4.5, See 5), é uma ferramenta adequada ao uso da referida técnica.

As Árvores de Decisão são representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados (GARCIA, 2000). Uma Árvore de Decisão utiliza a estratégia chamada **dividir-para-conquistar**, ou seja, um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (GAMA, 2000). A capacidade de discriminação de uma Árvore de Decisão advém das características de divisão do espaço definido pelos atributos em subespaços e da associação de uma classe a cada subespaço.

Segundo Garcia (2000), as Árvores de Decisão consistem de: nodos (nós), que representam os atributos, e de arcos (ramos), provenientes desses nodos e que recebem os valores possíveis para esses atributos (cada ramo descendente corresponde a um possível valor desse atributo). Nas árvores existem nodos folha (folha da árvore), que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está associada a uma classe. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

A figura 2 apresenta um exemplo de Árvore de Decisão, na qual constam dados que relatam as condições para uma pessoa receber um empréstimo. Nesse caso existem duas possíveis classes: Sim (receber empréstimo) e Não (não receber empréstimo). Os atributos são montante, salário e conta. O atributo Montante pode assumir os Valores de Médio, Alto ou Baixo; o atributo Salário pode assumir Valor Baixo ou Valor Alto; e o atributo Conta pode ser Sim ou Não. Alguns dados são exemplos da classe Sim, ou seja, os requisitos exigidos por um banco a uma pessoa para a concessão de um empréstimo são satisfatoriamente preenchidos. Outros são da classe Não, isto é, os requisitos exigidos não são plenamente satisfeitos. A classificação, nesse caso, resulta numa estrutura de árvore, que pode ser usada para todos os objetos do conjunto (BRADZIL, 1999).

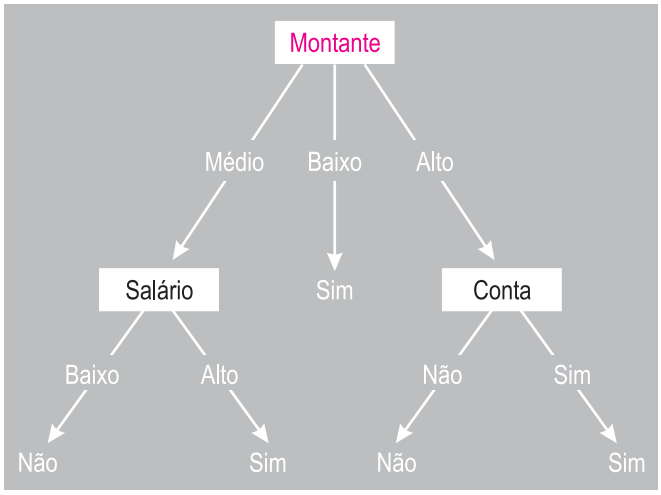

Figura 2: Exemplo de uma Árvore de Decisão

É possível derivar regras de uma Árvore de Decisão, com o intuito de facilitar a leitura e a compreensão por parte do usuário. Assim, as Árvores de Decisão podem ser representadas como conjuntos de regras do tipo **se-então** (*if-then*). As regras são escritas considerando o trajeto do nó raiz até uma folha da árvore. Árvores de Decisão e Regras de Classificação são métodos geralmente utilizados em conjunto. Devido ao fato de as Árvores de Decisão tenderem a crescer muito, como mostram algumas aplicações, elas são muitas vezes substituídas pelas regras. Isso acontece em virtude dessas últimas poderem ser facilmente modularizadas. Uma regra pode ser compreendida sem que haja a necessidade de referenciar-se a outras (INGARGIOLA, 1996).

Com base na Árvore de Decisão apresentada na figura 2, pode-se exemplificar a derivação de regras. Dois exemplos de regras obtidas a partir dessa árvore são mostrados a seguir:

- **Se** montante = médio **e** salário = baixo **então** classe = não.
- **Se** montante = médio **e** salário = alto **então** classe = sim.

Muitos são os algoritmos de classificação que utilizam a representação sob o formato de Árvores de Decisão (WITTEN e FRANK, 2000). O algoritmo ID3, cuja criação se baseou em sistemas de inferência e em conceitos de sistemas de aprendizagem, foi um dos primeiros de Árvore de Decisão. Logo após, foram elaborados diversos outros algoritmos, sendo os mais conhecidos o C4.5, o CART (*Classification and Regression Trees*) e o CHAID (*Chi Square Automatic Interaction Detection*), além de outros (GARCIA, 2000).

A utilização de Árvores de Decisão apresenta as seguintes vantagens: não assumem nenhuma distribuição particular para os dados; as características ou os atributos podem ser categóricos (qualitativos) ou numéricos (quantitativos); pode-se construir modelos para qualquer função desde que o número de exemplos de treinamento seja suficiente; elevado grau de compreensão.

Após a construção de uma Árvore de Decisão, é importante avaliá-la por meio da utilização de dados que não tenham sido usados no treinamento. Essa estratégia permite estimar como a árvore generaliza os dados e adapta-se a novas situações, além de determinar a proporção de erros e acertos ocorridos na construção da árvore (BRADZIL, 1999).

4.2. Redes neurais

Dentre os diversos modelos de Redes Neurais existentes, decidiu-se utilizar as redes de múltiplas camadas (*MultiLayer Perceptron* — MLP), que são modelos de redes que apresentam uma ou mais camadas de neurônios entre as camadas de entrada de dados e de saída dos resultados, chamadas camadas intermediárias (KRÖSE e VAN DER SMAGT, 1993). Esse tipo de Rede Neural artificial é o modelo mais utilizado atualmente, sendo em geral treinado através do algoritmo de Retropropagação (*Backpropagation).*

Nessas redes, cada camada tem uma função específica. A camada de saída recebe os estímulos da camada intermediária e gera a resposta final. As camadas intermediárias funcionam como extratoras de características, sendo seus pesos uma codificação de características apresentadas nos padrões de entrada e permitem que a rede crie sua própria representação, mais rica e complexa, do problema (CARVALHO, 2000). Essas camadas intermediárias são unidades que não interagem diretamente com o ambiente, daí sua denominação.

Se existirem conexões apropriadas entre as unidades de entrada e um conjunto suficientemente grande de unidades intermediárias, pode-se sempre encontrar a representação que irá produzir o mapeamento correto entre a entrada de dados e a saída dos resultados (classificação) por meio das unidades intermediárias.

O algoritmo de Retropropagação (*Backpropagation*) é o método mais utilizado para o treinamento de redes neurais e enquadra-se na categoria de aprendizagem supervisionada, ou seja, os resultados desejados são conhecidos preliminarmente e o treinamento da Rede Neural é feito com o objetivo de **fazê-la aprender** a obter esses resultados. Esse algoritmo geralmente é aplicado a redes com múltiplas camadas do tipo em avanço (*feed-forward).* Durante o treinamento com o referido algoritmo, a rede opera em uma seqüência de dois passos. No primeiro passo um padrão (de um total de *m* padrões; nesse caso *m* = 339 = número total de empresas) é apresentado à camada de entrada da rede. A atividade resultante flui através da rede, camada por camada, até que a resposta seja produzida pela camada de saída. No segundo passo, a saída obtida é comparada à saída desejada para esse padrão particular e o erro (diferença entre o valor desejado e o valor obtido) é calculado. O erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo ajustados conforme o erro é retropropagado (CARVALHO, 2000).

Nesse segundo passo, geralmente se utiliza o método do gradiente de uma função, que calcula as derivadas parciais do erro com respeito ao vetor peso, W, de um certo vetor de entradas, X (padrão). Esse método aponta a quantidade de erro do resultado (saída), de modo a corrigir os pesos para que se produza a saída desejada diante da respectiva entrada. Esses dois passos para o treinamento da rede neural são chamados de propagação avante (*forward)* e propagação de retorno (*backward*), respectivamente.

O treinamento das redes de várias camadas pelo algoritmo de retropropagação pode demandar muitas iterações do conjunto de treinamento, resultando em um tempo de treinamento longo. Se for encontrado um mínimo local, o erro para o conjunto de treinamento estabiliza-se, estacionando em um valor muitas vezes maior do que o desejado. Por isso, deve-se fazer uma série de testes computacionais, iniciando o treinamento de diferentes pontos iniciais. Deve-se também proceder dessa forma para diferentes arquiteturas da Rede Neural, ou seja, para diversos tamanhos da camada oculta.

5. IMPLEMENTAÇÃO COMPUTACIONAL E OBTENÇÃO DOS RESULTADOS

Nesta seção é apresentada a implementação computacional das duas técnicas abordadas, Árvores de Decisão e Redes Neurais, ao problema descrito na seção 2, bem como a comparação entre elas no que diz respeito a seus desempenhos na obtenção dos resultados.

5.1. Árvores de decisão

Na implementação da técnica Árvores de Decisão optou-se por utilizar o *software* computacional WEKA (*Waikato Environment for Knowledge Analysis*), por sua praticidade de utilização e por ser um *software* de domínio público disponí-

vel em: <http://www.cs.waikato.ac.nz/ml/weka/>. O *software* WEKA é formado por um conjunto de diversas técnicas para resolver problemas concretos de Mineração de Dados. O WEKA está implementado em linguagem Java e foi desenvolvido na Universidade de Waikato, na Nova Zelândia, em 1999.

O *software* WEKA utiliza o padrão ARFF para seus arquivos de entrada, independentemente do algoritmo utilizado. Esse padrão é usado para representar uma série de dados que consistem em exemplos independentes. A especificação dos atributos em linhas de ARFF permite que os dados sejam verificados, de forma automática, quanto à sua consistência pelos programas que lêem as linhas de ARFF (WITTEN e FRANK, 2000).

Na implementação da técnica de Árvores de Decisão, foi utilizada a base de dados deste trabalho composta por informações de 339 empresas (266 adimplentes e 73 inadimplentes), considerando-se para cada uma delas as 24 características apresentadas no quadro da seção 2 deste artigo, empregando na execução o algoritmo de classificação J48 (C4.5 *release* 8).

Foram realizados oito conjuntos de testes no total. O primeiro teste contemplou informações de todas as 339 empresas. Nos demais, os dados foram separados em dois conjuntos: um de treinamento com os dados de 306 empresas, composto por 241 empresas adimplentes e 65 inadimplentes, e outro de testes com informações de 33 empresas, composto por 25 empresas adimplentes e 8 inadimplentes. Exceto no primeiro caso, em cada um dos testes realizados, os conjuntos foram gerados de forma aleatória, a fim de evitar qualquer tipo de indução de resultados. Os resultados obtidos são apresentados da tabela 1. Com base nos dados apresentados nessa tabela, observa-se que o Teste 5 apresentou o melhor resultado, não só no contexto envolvendo o conjunto utilizado para gerar a Árvore de Decisão, como também no conjunto utilizado para testá-la. A árvore gerada pelo Teste 5 apresentou 282 padrões (ou seja, empresas) classificados corretamente e 24 registros classificados de forma incorreta, correspondendo a uma taxa de acerto de 92,16%. Dos 282 registros classificados corretamente, 235 referem-se a empresas adimplentes e 47 a empresas inadimplentes. Dos 24 registros classificados incorretamente, 6 são de empresas adimplentes e 18 de empresas inadimplentes.

Na Árvore de Decisão gerada no Teste 5 existem 40 folhas, ou seja, podem-se extrair 40 regras do tipo **se-então**, algumas das quais relacionadas a seguir:

- **Se** tempo de conta > 25 meses, **então** adimplente.
- **Se** tempo de conta $\leq$ 25 meses **e** sócios não têm restrições **e** a empresa possui risco A, **então** adimplente.
- **Se** tempo de conta $\leq$ 25 meses **e** sócios não têm restrições **e** a empresa possui risco B **e** sócios não tiveram restrições baixadas nos últimos cinco anos **e** a sociedade é entre cônjuges, **então** adimplente.
- **Se** tempo de conta $\leq$ 25 meses **e** sócios não têm restrições **e** a empresa possui risco C **e** a empresa teve restrições baixadas nos últimos cinco anos, **então** inadimplente.
- **Se** tempo de conta $\leq$ 25 meses **e** sócios não têm restrições **e** a empresa possui risco C **e** a empresa não teve restrições baixadas nos últimos cinco anos **e** principais clientes são pessoas físicas **e** não possui seguro empresarial, **então** inadimplente.

5.2. Redes neurais

Para a implementação da técnica de Redes Neurais ao problema, foi utilizado o *software* MATLAB (*Neural Networks Toolbox*). Na utilização da técnica de Redes Neurais procedeu-se de maneira análoga à utilizada na técnica de Árvores

Tabela 1

Resultados Obtidos com a Técnica de Árvores de Decisão

Testes	Árvores de Decisão					
	Conjunto de Treinamento			Conjunto de Testes		
	Adimplentes	Inadimplentes	Erro(%)	Adimplentes	Inadimplentes	Erro(%)
1	12/266	23/73	10,32	—	—	—
2	0/241	65/65	21,25	0/25	7/8	21,21
3	7/241	25/65	10,45	6/25	4/8	30,30
4	15/241	30/65	14,71	8/25	4/8	36,36
5	6/241	18/65	7,84	6/25	2/8	24,24
6	12/241	16/65	9,15	9/25	3/8	36,36
7	2/241	22/65	8,05	4/25	4/8	24,24
8	5/241	26/65	10,20	5/25	3/8	24,24
Média	—	—	11,49	—	—	28,13

de Decisão, ou seja, foram realizados oito conjuntos de testes no total. O primeiro contemplou informações de todas as 339 empresas. Nos demais, os dados foram separados em dois conjuntos: um de treinamento com dados de 306 empresas, composto por 241 empresas adimplentes e 65 inadimplentes, e outro de testes com informações de 33 empresas, composto por 25 empresas adimplentes e oito inadimplentes. Os conjuntos de treinamento e de testes foram os mesmos conjuntos utilizados quando do uso da técnica Árvores de Decisão.

Os treinamentos foram feitos por meio de uma rede de múltiplas camadas, usando o algoritmo de Retropropagação padrão, variando os seguintes parâmetros:

- quantidade de iterações (ou ciclos) — em cada conjunto de testes, a Rede Neural foi treinada com 100, 1.000, 2.000, 4.000, 6.000, 8.000 e 10.000 iterações;
- quantidade de neurônios intermediários da rede — em cada teste realizado, a Rede Neural foi treinada primeiramente sem a camada intermediária e, a seguir, utilizando 2, 4, 6, 8 e 10 neurônios na camada intermediária. Para cada teste realizado, foi utilizado um conjunto aleatório de pesos iniciais, num total de 48 conjuntos;
- em todos os testes, foi utilizada a taxa de aprendizagem igual a 0,01 e optou-se por não utilizar a taxa *momentum*. A quantidade de neurônios na camada de entrada é igual ao número de variáveis utilizadas, no caso deste artigo igual a 24, e a quantidade de neurônios na camada de saída é igual a 1.

Os resultados encontrados em cada conjunto de testes constam da tabela 2.

Os resultados apresentados na tabela 2 foram obtidos após 10.000 iterações em cada conjunto de treinamento. Analisando-os, observa-se que o melhor resultado obtido com a técnica de Redes Neurais foi o encontrado no Teste 6, que apresenta, na fase de treinamento, uma taxa de erro de 2,28% e, na fase de testes, de 3,03%.

Os resultados obtidos, do ponto de vista do percentual de erros apresentados com o uso das duas técnicas, encontram-se sintetizados na tabela 3.

Tabela 3

Média dos Erros nas Técnicas de Árvores de Decisão e Redes Neurais

Conjunto	Erro Médio de Classificação	
	Árvores de Decisão %	Redes Neurais %
Treinamento	11,49	4,09
Testes	28,13	9,96

O desempenho da técnica de Redes Neurais foi melhor do que o apresentado pela técnica de Árvores de Decisão em relação à taxa de classificação correta, conforme pode ser verificado na tabela 3. Do ponto de vista do usuário (gerente bancário ou analista de crédito), porém, sempre há vantagens no uso da técnica de Árvores de Decisão, no sentido de que ela apresenta resultados (regras de decisão) de fácil compreensão, detalhando quais das informações sobre as empresas analisadas foram mais relevantes na classificação. Dessa forma, o usuário pode verificar se os resultados fornecidos pela técnica estão de acordo com a sua experiência. Com o uso das técnicas aqui apresentadas, é possível proceder à análise de novas propostas de concessão de crédito, com uma margem maior de segurança, fornecendo, assim, uma ferramenta auxiliar para o usuário.

Vale salientar que os dados a serem apresentados para a Rede Neural precisam ser transformados em valores numéricos, como descrito na seção 2. Já na técnica de Árvores de

Tabela 2

Resultados Obtidos com a Técnica de Redes Neurais

Testes	Quantidade de Neurônios Camada Escondida	Redes Neurais					
		Conjunto de Treinamento			Conjunto de Testes		
		Adimplentes	Inadimplentes	Erro(%)	Adimplentes	Inadimplentes	Erro(%)
1	8	13/270	4/69	5,01	—	—	—
2	8	4/241	7/65	3,59	2/25	1/8	9,09
3	10	8/241	6/65	4,57	2/25	1/8	9,09
4	10	6/241	7/65	4,24	2/25	1/8	9,09
5	8	2/241	12/65	4,57	1/25	3/8	12,12
6	10	1/241	6/65	2,28	0/25	1/8	3,03
7	8	1/241	10/65	3,59	1/25	3/8	12,12
8	8	7/241	8/65	4,90	2/25	3/8	15,15
Média	—	—	—	4,09	—	—	9,96

Decisão, é possível a entrada de dados com seus valores no formato original (quantitativos e qualitativos).

6. CONCLUSÕES

As técnicas de Mineração de Dados apresentadas e testadas neste trabalho, Árvores de Decisão e Redes Neurais, no contexto de KDD, mostraram ser ferramentas de grande valia para os analistas de crédito bancário. Desse modo, utilizando as informações cadastrais de empresas, os analistas têm condições de diagnosticar novas empresas em relação ao merecimento de crédito. Neste artigo, os resultados fornecidos pelos dois métodos utilizados comprovaram sua eficiência na classificação das empresas como adimplentes ou inadimplentes.

Convém ressaltar que, mesmo com uma sinalização totalmente favorável à concessão de crédito a um novo cliente, ele pode vir a tornar-se um cliente inadimplente, visto que a economia ainda não está totalmente estabilizada. Outros fatores, tais como um sinistro (incêndio, roubo ou outro), podem interferir no comportamento da empresa em face dos compromissos assumidos.

A tarefa de conceder ou não crédito é e sempre será difícil. De qualquer modo, ferramentas quantitativas como as apresentadas, aliadas à experiência do analista de crédito, são imprescindíveis. Redes Neurais e Árvores de Decisão são ferramentas que podem ser utilizadas pelo especialista para auxiliá-lo na tomada de decisões. Contudo, dificilmente elas poderão, por si sós, substituir a figura do especialista no processo de análise de crédito.

A fim de evitar a inadimplência, após a concessão do crédito devem ser adotadas medidas de controle de seu retorno nos prazos previstos, tarefa tão importante quanto a decisão de conceder ou não o crédito.◆

REFERÊNCIAS BIBLIOGRÁFICAS

BRADZIL, P. B. *Construção de modelos de decisão a partir de dados*. 1999. Disponível em: <http://www.nacc.up.pt/~pbrazdil/Ensino/ML/ModDecis.html>. Acesso em: 21 jul. 2002.

CARVALHO, A.P. de L.F. de. *Redes neurais artificiais*. 2000. Disponível em: <http://www.icmc.sc.usp.br/~andre/neural1.html>. Acesso em: 14 jul. 2002.

DEPARTAMENTO DE INFORMÁTICA (DIN). Universidade Estadual de Maringá (UEM). Grupo de Sistemas Inteligentes (GSI). *Mineração de dados*. 1998. Disponível em: <http://www.din.uem.br/ia/mineracao/tecnologia/ferramentas.html>. Acesso em: 14 jul. 2002.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. *Advances in knowledge discovery & data mining*. Menlo Park, CA, USA: AAAI/MIT, 1996.

FREITAS, A.A. Uma introdução a data mining. *Informática Brasileira em Análise*, Recife, Centro de Estudos e Sistemas Avançados do Recife (CESAR), ano II, n.32, p.34, maio/jun. 2000.

GAMA, J. *Árvores de decisão*. 2000. Disponível em: <http://www.liacc.up.pt/~jgama/Mestrado/ECD1/Arvores.html>. Acesso em: 14 ago. 2002.

GARCIA, S.C. *O uso de árvores de decisão na descoberta de conhecimento na área da saúde*. In: SEMANA ACADÊMICA, 2000. Rio Grande do Sul: Universidade Federal do Rio Grande do Sul, 2000.

INGARGIOLA, G. *Building classification models*: ID3 and C4.5. 1996. Disponível em: <http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html>. Acesso em: 24 ago. 2002.

INMON, W.H. *Como construir o data warehouse*. Rio de Janeiro: Campus, 1997.

KRÖSE, B.J.A.; VAN DER SMAGT, P.P. *An introduction to neural networks.* Amsterdam: University of Amsterdam, 1993.

MANNILA, H. *Data mining*: machine learning, statistics, and databases. In: INTERNATIONAL CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANANGEMENT, Stockholm, 1996. p.1-8.

NIMER, F.; SPANDRI, L.C. Obtendo vantagens competitiva com o uso de data mining. *Developers Magazine*, *Ano* 2, n.18, p.32, Feb. 1998.

POSSAS, B.A.V.; CARVALHO, M.L.B. de; REZENDE, R.S.F.; MEIRA JR., W. *Data mining*: técnicas para exploração de dados. Belo Horizonte: Universidade Federal de Minas Gerais, 1998.

QUINLAN, J.C. *C4.5*: programs for machine learning. San Mateo: Morgan Kaufmann, 1993. 302p.

STEINER, M.T.A.; CARNIERI, C.; KOPITTKE, B.H.; STEINER NETO, P.J. Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário. *Revista de Administração da Universidade de São Paulo* (RAUSP), São Paulo, v.34, n.3, p.56-67, jul./set. 1999.

WITTEN, I.H.; FRANK, E. *Data mining*: pratical machine learning tools and techniques with Java implementations. San Francisco, California: Morgan Kaufmann, 2000.

ABSTRACT

Bank credit analysis with the use of neural networks and decision trees: a simple data mining application

In the bank credit area, the possession and use of tools that can help identify and forecast which clients will be "good or bad" credit takers could be a key factor and result in greater competitive advantage. There is a lot of "hidden" knowledge in the immense quantity of data that is available in companies' databases. With the use of data mining methodologies, one can transform these data into valuable information, which would help in decision processes. In this paper, we analyze the historical data of 339 clients (legal entities) from a bank agency by means of two data mining tools: neural networks and decision trees. These techniques allow the recognition of patterns, as well as the diagnosis of new cases. Results were quite satisfactory and showed that for this specific problem, neural networks had a smaller error percentage.

***Uniterms*:** KDD, data mining, decision trees, neural networks, bank credit.

RESUMEN

Análisis de crédito bancario por medio de redes neuronales y árboles de decisión: una aplicación simple de *data mining*

En el área de crédito bancario, el dominio y el uso de herramientas que ayuden en la tarea de clasificación de clientes en probables pagadores o deudores con relación a la toma de crédito pueden convertirse en un factor clave, resultando en una gran ventaja competitiva. Dentro de la inmensa cantidad de datos disponibles en las bases de datos de las empresas hay muchos conocimientos útiles e importantes que están **escondidos**. Con la metodología de *Data Mining*, estos datos se pueden convertir en informaciones valiosas que contribuirán con el proceso decisorio. En este trabajo se analizan registros históricos de 339 clientes (personas jurídicas) de una agencia bancaria, por medio de dos de las herramientas de *Data Mining*: Redes Neuronales y Árboles de Decisión. Dichas técnicas permiten llevar a cabo el reconocimiento de patrones y también clasificar nuevos casos. Los resultados fueron bastante satisfactorios y mostraron que, para ese problema específico, las Redes Neuronales presentaron una tasa de clasificación correcta mayor que la de los Árboles de Decisión.

***Palabras clave*:** KDD, *data mining*, árboles de decisión, redes neuronales, crédito bancario.