ELIANE PREZEPIORSKI LEMOS

ANÁLISE DE CRÉDITO BANCÁRIO COM O USO DE *DATA MINING*: REDES NEURAIS E ÁRVORES DE DECISÃO

CURITIBA
2003

ELIANE PREZEPIORSKI LEMOS

ANÁLISE DE CRÉDITO BANCÁRIO COM O USO DE *DATA MINING*: REDES NEURAIS E ÁRVORES DE DECISÃO

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências, do Programa de Pós-Graduação em Métodos Numéricos em Engenharia, na Área de Concentração em Programação Matemática, Setor de Tecnologia, Departamento de Construção Civil e Setor de Ciências Exatas, Departamento de Matemática da Universidade Federal do Paraná.

Orientadora: Profa. Dra. Maria Teresinha Arns Steiner
Co-orientador: Prof. Dr. Alex Alves Freitas

CURITIBA

2003

TERMO DE APROVAÇÃO

ELIANE PREZEPIORSKI LEMOS

ANÁLISE DE CRÉDITO BANCÁRIO COM O USO DE
DATA MINING: REDES NEURAIS E ÁRVORES DE DESCISÃO

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Ciências, do Programa de Pós-Graduação em Métodos Numéricos em Engenharia, na Área de Concentração em Programação Matemática, Setor de Tecnologia, Departamento de Construção Civil e Setor de Ciências Exatas, Departamento de Matemática da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientadora:

Profa. Dra. Maria Teresinha Arns Steiner
Departamento de Matemática, UFPR

Prof. Dr. Osmar Ambrósio de Souza
Departamento de Matemática, UNICENTRO

Prof. Dr. Ricardo Mendes Júnior
Departamento de Construção Civil, UFPR

Prof. Dr. Júlio Cesar Nievola
Departamento de Ciência da Computação, PUC-PR

Curitiba, 09 de julho de 2003

Dedico à minha mãe Lydia (in memorian), meu exemplo de vida.

AGRADECIMENTOS

Aos meus pais, pela minha vida e por me ensinarem a valorizar a educação.

Ao meu marido Luiz Antonio, que me incentivou na realização desse projeto e que não me deixou fraquejar nos momentos difíceis.

Às minhas filhas, Alessandra e Milena, que apesar da pouca idade, souberam entender e aceitar todas as vezes em que não pude lhes dar a atenção que desejavam.

À minha irmã, Adriane, minha melhor amiga, que sempre esteve do meu lado, em todos os momentos da minha vida, compartilhando e torcendo pelo meu sucesso.

Ao meu cunhado Leomar, pelos diversos auxílios.

À professora Maria Teresinha Arns Steiner, pela orientação ao longo desta dissertação, pela prestividade, enorme paciência, pelo incentivo e confiança e, principalmente, pela amizade.

Ao professor Alex Alves Freitas, meu co-orientador, pelos valiosos conhecimentos transmitidos.

Ao professor Celso Carnieri, a quem aprendi a admirar e querer bem.

A todos os professores do curso, pelos ensinamentos.

Aos colegas de curso, pelo companheirismo.

Às colegas Karina, Luciene, Eliana, Elisane e Margarete que ao longo dessa jornada, deixaram de ser simplesmente colegas para serem minhas amigas. Obrigada a todas vocês por tantas e tantas vezes me estenderem a mão nessa caminhada.

À UNICENTRO, principalmente ao Departamento de Matemática, pela oportunidade de cursar esse Mestrado.

Ao professor Osmar Ambrósio de Souza, pelo incentivo e pela grande ajuda com o programa para realizar os testes de Redes Neurais.

Ao Banco do Brasil S.A., agência de Guarapuava (PR), por disponibilizar os dados utilizados neste trabalho.

À todos aqueles que, de uma forma ou de outra, colaboraram para a realização deste trabalho.

E, principalmente, agradeço à Deus, meu grande companheiro, pela força e pelo amparo nesta jornada.

"Todo homem, por natureza, deseja o Conhecimento".

Aristóteles (384-322 a.C.)

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O mundo dos negócios está mais competitivo do que nunca. Especificamente no caso de crédito bancário, possuir e utilizar ferramentas que possam auxiliar na tarefa de reconhecer e prever quais clientes serão "bons ou maus" tomadores de crédito, pode se tornar um fator chave, resultando numa grande vantagem competitiva. Existe muito conhecimento escondido na imensa quantidade de dados disponíveis nos bancos de dados das empresas. Com a metodologia de *Data Mining,* pode-se transformar esses dados em informações valiosas para auxiliar no processo decisório. Neste trabalho estão sendo analisados registros históricos de 339 clientes (pessoas jurídicas) de uma agência bancária, através de duas das ferramentas de *Data Mining:* Redes Neurais e Árvores de Decisão. Estas técnicas permitem fazer o reconhecimento de padrões e também diagnosticar novos casos. A idéia central deste trabalho é, portanto, utilizando as técnicas de Redes Neurais e Árvores de Decisão através do uso dos *softwares* MATLAB-*Neural Networks Toolbox* e WEKA-*Waikato Environment for Knowledge Analysis,* respectivamente, auxiliar na tomada de decisão sobre conceder ou não crédito bancário a um novo cliente. Os resultados foram bastante satisfatórios, mostrando que, para este problema específico, as Redes Neurais apresentaram um percentual menor de erros.

ABSTRACT

The business world is more competitive than ever. Specifically in the case of bank credit, the use of tools that can assist in the task to recognize and to foresse which customers will be "good or bad" credit payers, can become a key factor, resulting in a great competitive advantage. There is hidden knowledge in the immense available amount data in the databases of companies. With the Data Mining methodology, these data can be transformed into valuable information to assist in the making decision process. In this work we analyzed historical registers of 339 customers of a bank agency, through two of the tools of Data Mining: Neural Networks and Decision Trees. These techniques allow to make the pattern recognition and also to diagnose new cases. The central idea of this work is therefore, using the techniques of Neural Networks and DecisionTrees through the use of MATLAB-Neural Networks Toolbox and WEKA-Waikato Environment for Knowledge Analysis softwares, respectively, to assist in the making decision on granting or not credit bank to a new customer. The results had been sufficiently satisfactory, showing that, for this specific problem, the Neural Nets had presented a lesser percentage of errors.

CAPÍTULO I

1. INTRODUÇÃO

O ambiente de negócios atual está mais competitivo do que nunca. Especificamente no caso de crédito bancário, possuir e utilizar ferramentas que possam auxiliar na tarefa de reconhecer e prever quais clientes serão "bons ou maus" tomadores de crédito, é um fator chave, resultando numa grande vantagem competitiva.

Com os recursos da informática, as empresas podem acumular uma imensa quantidade de dados das mais variadas fontes. São informações industriais, comerciais, relatórios de vendas, hábitos de ligações telefônicas, hábitos de compras, dentre outras.

A partir da década de 80, algumas empresas começaram a descobrir que, em meio a essa imensa quantidade de dados, podiam se esconder informações valiosas. Começaram a perceber que não bastava simplesmente armazenar dados, era preciso transformá-los em informação.

Um dado se transforma em informação quando ganha um significado para seu utilizador, caso contrário, continua sendo simplesmente um dado.

Então surgiu uma pergunta inevitável: como explorar todos esses dados?

Métodos estatísticos, desenvolvidos há décadas, vêm sendo exaustivamente utilizados, porém novas metodologias para extração de conhecimento estão sendo estudadas, dentre elas *Data Mining.*

Data Mining ou Mineração de Dados é uma nova metodologia para melhorar a qualidade e eficiência das decisões, quer sejam elas científicas ou de negócios.

Muitas companhias vêm obtendo altos retornos sobre seus investimentos em banco de dados e ferramentas analíticas. Um caso recente de descoberta literalmente preciosa com a mineração de dados foi relatado no *The Wall Street Journal.* O jornal relata que o cassino Harrah's, em Las Vegas, "minerou" o seu banco de dados (com o perfil de 16 milhões de clientes) e descobriu uma informação valiosíssima: os apostadores que gastam entre 100 e 500 dólares numa visita ao cassino correspondem a apenas 30% de toda a clientela, mas contribuem com 80% das receitas. Com estratégias agressivas de marketing para atrair esse filão mais rentável (são oferecidos almoços, shows e apostas grátis), o cassino afirma ter dobrado seu faturamento no ano seguinte (GUIZZO, 2000).

As empresas, de um modo geral, buscam mecanismos, ferramentas que possam agregar algum diferencial mercadológico e proporcionar uma maior rentabilidade e segurança nas suas relações comerciais. No caso dos bancos, a realidade também é a mesma. Em função do processo de estabilização da moeda brasileira, os bancos brasileiros têm sido obrigados a passar por um processo rigoroso de adaptação: a forte e rápida redução dos ganhos inflacionários forçou os bancos a aumentarem seus volumes de crédito (ALMEIDA; SIQUEIRA, 1997).

Segundo STEINER et al. (1999), "a correta decisão de crédito é essencial para a sobrevivência das empresas bancárias. 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 é,

portanto, analisar uma proposta de negócio e comparar o 'custo de conceder' com o 'custo de negar' a operação".

Surge assim, a necessidade de se ter ferramentas que possam auxiliar nas decisões de conceder ou não o crédito.

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

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

1.1. OBJETIVOS DO TRABALHO

O objetivo principal deste trabalho é gerar classificadores para que através destes se possa fazer a classificação de novas empresas como adimplentes ou inadimplentes, utilizando técnicas de *Data Mining*.

Estas técnicas permitem fazer o reconhecimento de padrões e também sua utilização para fazer futuros diagnósticos.

Dentre as diversas técnicas de *Data Mining*, tais como: Análise de *Cluster*, Árvores de Decisão, Redes Neurais, Indução de Regras, Algoritmos Genéticos, Aprendizado baseado em casos, optou-se pela utilização de duas delas: Redes Neurais e Árvores de Decisão.

A intenção de se utilizar duas técnicas é fazer a comparação dos resultados obtidos em cada uma delas, verificando-se assim qual oferece a menor porcentagem de erros para o contexto do presente trabalho, ou seja, na classificação de novas empresas.

1.2. LIMITAÇÕES DO TRABALHO

Esse trabalho não contempla dados de empresas de médio e grande porte, tendo em vista que a análise de crédito para empresas desses portes não é feita no âmbito das agências do Banco do Brasil e sim em um órgão interno do Banco, denominado Divisão de Análise de Crédito, onde especialistas para efetuarem a análise utilizam, além dos dados cadastrais, dados contábeis (balanços e balancetes) referentes aos três últimos exercícios contábeis. Assim, os dados utilizados e trabalhados limitam-se a micro e pequenas empresas, cuja análise é feita essencialmente através dos dados cadastrais, contidos no ANEXO I.

1.3. REVISÃO DA LITERATURA - TRABALHOS RELACIONADOS

Neste item é feita uma revisão bibliográfica, com um rápido relato de outros trabalhos desenvolvidos com a utilização das técnicas de Redes Neurais e Árvores de Decisão.

As técnicas de Redes Neurais e Árvores de Decisão vêm sendo amplamente utilizadas nas últimas décadas, com aplicações em diversos campos, tais como: área da saúde, marketing, na área de crédito, na detecção de fraudes e várias outras.

ALMEIDA e DUMONTIER (1996), publicaram um trabalho onde apresentam uma abordagem estruturada da exploração de Redes Neurais para avaliação de riscos de inadimplência, avaliando o setor de transporte de carga rodoviário francês. O desempenho das Redes Neurais foi comparado com o Método da Regressão Logística. O desempenho das Redes Neurais não foi significativamente superior ao desempenho do método estatístico.

TERRA e PEREIRA (1999), escreveram um artigo onde apresentaram uma metodologia de solução do problema de programação da produção, através da utilização combinada da técnica de simulação de sistemas com a abordagem por Redes Neurais Artificiais. Nesse trabalho foram apresentados os principais aspectos relativos à programação da produção, os principais conceitos sobre Redes Neurais Artificiais e a justificativa da utilização dessa ferramenta na programação da produção.

CARLSON e TAVARES (1995), apresentaram um artigo propondo a representação de inequações lineares com variáveis 0-1 através de Redes Neurais. Apresentaram nesse trabalho uma extensão da regra devida a TAGLIARINI et al.[1] para transformar sistematicamente inequações com variáveis 0-1 e coeficientes lineares em função energia de uma Rede Neural de Hopfield. Foi criada uma Rede Neural para representar um problema pequeno, que mostrou possuir configurações estáveis que correspondem a soluções do problema original. Ou seja, a rede foi capaz de incorporar, em seus pesos e entradas, o relacionamento entre as restrições e o objetivo do problema.

---

[1] TAGLIARINI, G., J.F. CHRIST e E.W. PAGE. "Optimization Using Neural Networks". IEEE Transactions on Computers 40(12), p.1347-1358, 1991.

ALMEIDA (1995), publicou um artigo onde apresenta uma visão geral do potencial e do funcionamento do uso de Redes Neurais em administração, através de exemplos e ilustrações acessíveis ao leitor não-familiarizado com conceitos de informática.

ALMEIDA e SIQUEIRA (1997), fazem uma comparação entre Regressão Logística e Redes Neurais na previsão de falência de bancos brasileiros. Apesar das Redes Neurais não apresentarem resultados muito superiores aos obtidos pela Regressão Logística, apresentou um fator diferencial que foi o de considerar bancos que a Regressão Logística não pode classificar por falta de dados.

STEINER et al., (1999), utilizaram Sistemas Especialistas Probabilísticos e Redes Neurais na análise do crédito bancário para pessoas físicas, com o objetivo de prever o comportamento de futuros clientes com relação a adimplência ou não.

BOSIGNOLI e INFANTOSI (1999), publicaram um artigo onde estabelecem uma classificação automática do estado de sono ativo neonatal usando Redes Neurais Artificiais. O desempenho das Redes Neurais resultou em 95% de classificação correta, resultados estes, que quando comparados com a literatura sobre o assunto, indicam a potencialidade deste classificador de sono ativo.

FRANCO e MARTINS (1999), fazem uso do algoritmo de classificação de dados C4.5 e Redes Neurais no acasalamento de gado nelore.

ADAMOWICZ (2000), apresenta um trabalho para reconhecimento de padrões na análise econômico-financeira de empresas, a fim de discriminar empresas solventes de empresas insolventes. Utilizou dois métodos: um estatístico, Análise Discriminante Linear de Fisher e um na área de Inteligência Artificial, Redes Neurais.

GARCIA (2000), faz uso de Árvores de Decisão na descoberta de conhecimento na área da saúde.

NOGUEZ (2000), propôs um estudo sobre a implementação de Árvores de Decisão com Múltiplos Classificadores.

CUROTTO (2002), apresenta uma estratégia de *Data Mining* baseada em indução incremental de Árvores de Decisão. São abordados problemas de classificação e de regressão.

FREITAS JÚNIOR et al. (2001), utilizaram uma base de dados do quadro docente da Universidade Estadual de Maringá-UEM, no qual foram aplicadas técnicas de *Data Mining* . Obtiveram como resultados uma série de regras de associação, as quais poderão servir como indicativos para a direção da UEM no tocante ao auxílio ao processo decisório, principalmente no que se refere a investimentos em infra-estrutura e programas de capacitação de seu corpo docente.

1.4. ESTRUTURA DO TRABALHO

Este trabalho foi dividido em mais cinco capítulos, além desta introdução.

O capítulo dois apresenta a descrição do problema abordado neste trabalho.

No capítulo três são apresentados conceitos que envolvem o processo de Mineração de Dados e algumas de suas técnicas.

O capítulo quatro descreve em detalhes as técnicas aplicadas neste trabalho: Redes Neurais e Árvores de Decisão.

No quinto capítulo é feita a aplicação das técnicas ao problema e é efetuada a análise dos resultados.

Finalmente, no capítulo seis são apresentadas algumas conclusões obtidas das análises realizadas no capítulo anterior, além de sugestões para realizações de futuros trabalhos.

CAPÍTULO II

2. DESCRIÇÃO DO PROBLEMA

2.1. INTRODUÇÃO

Os dados utilizados neste trabalho são reais e foram obtidos junto ao Banco do Brasil S.A., agência Guarapuava (PR), 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 segmento de mercado, colocando à disposição linhas de crédito para Capital de Giro como também para Investimentos. Outra informação importante é que 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 ANC - Análise de Crédito. É através 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.

Este trabalho procura analisar dados históricos de 339 clientes pessoa jurídica da agência de Guarapuava (PR), do Banco do Brasil S.A.

Através desta análise pretende-se verificar como esses dados históricos podem auxiliar no processo decisório da concessão de crédito.

2.2. OBTENÇÃO DOS DADOS

Os dados utilizados neste trabalho foram extraídos do banco de dados do Banco do Brasil S.A. - agência Guarapuava (PR), através de um formulário criado especialmente para esse fim, de acordo com o ANEXO I deste trabalho. Esses dados retratam o histórico das empresas com posição em junho de 2002. Ressalta-se assim, que não foram consideradas possíveis alterações ocorridas no histórico dessas empresas após junho de 2002, tendo em vista dificuldades na obtenção da atualização desses dados junto ao Banco do Brasil.

Os critérios para avaliação do crédito, portanto, são baseados nos referidos dados históricos, obtidos de 339 empresas, das quais 266 são adimplentes e 73 são inadimplentes, através de 29 informações sobre cada uma delas, que se encontram disponíveis no ANEXO II deste trabalho.

Assim, são essas informações que constituirão o banco de dados que será utilizado pelas técnicas aqui abordadas.

As 29 informações que constituirão as variáveis do problema que está sendo analisado, são apresentadas a seguir.

A. Existência de restrições em nome da empresa:

1 = SIM 2 = NÃO

B. Existência de restrições baixadas nos últimos 5 anos, em nome da empresa:

1 = SIM 2 = NÃO

C. Tempo de conta no Banco do Brasil (BB):

Esta informação foi coletada considerando a quantidade de anos e meses que a empresa possui conta no BB. Para o banco de dados essa informação foi traduzida para um valor numérico correspondente ao número total de meses.

Exemplo: Dado coletado = 1 ano e 5 meses

Dado atribuído = 17 (1 ano = 12 e 5 meses = 5 $\Rightarrow$ 17)

D. Setor de Atividade:

1 = COMÉRCIO 2 = INDÚSTRIA 3 = SERVIÇOS

E. Tempo de Atividade:

1 = mais de 9 anos

2 = 6 a 9 anos

3 = 3 a 5 anos

4 = 1 a 2 anos

5 = menos de 1 ano

F. Número de Funcionários: valor numérico

G. Sede da Empresa:

1 = PRÓPRIO 2 = ALUGADO 3 = CEDIDO

H. **Bairro:**

Informação coletada foi o nome do bairro. Dado correspondente atribuído no banco de dados foi:

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 BB:**

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 BB:

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 5 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:

1 = A

2 = B

3 = C

4 = D

5 = E

A variável "X" - risco atribuído pelo Banco - é um conceito definido pelo aplicativo ANC através do qual são estipuladas as garantias mínimas exigidas nas operações de crédito. Na escala, "A" é o melhor e "E" é o pior conceito.

Y. Resultado:

1 = ADIMPLENTE 2 = INADIMPLENTE

Z. Concentração das vendas nos 5 principais clientes:

1 = menos de 20% 2 = mais de 20%

AA. Crédito junto a Fornecedores:

1 = SIM 2 = NÃO

BB. Conceito na Praça:

1 = SIM 2 = NÃO

CC. Pontualidade:

1 = PONTUAL 2 = PAGA COM PEQUENOS ATRASOS

As variáveis Z, AA, BB e CC foram descartadas da base dados, uma vez que as respostas de todas as empresas pesquisadas para esses itens eram as mesmas, ou seja:

- **Z. Concentração das vendas nos 5 principais clientes:** todas as respostas coletadas foram iguais a menos de 20%.
- **AA. Crédito junto a Fornecedores:** todas respostas iguais a SIM.
- **BB. Conceito na Praça:** todas respostas iguais a SIM.
- **CC. Pontualidade:** todas as respostas iguais a PONTUAL.

CAPÍTULO III

3. KDD E *DATA MINING*

3.1. INTRODUÇÃO

Nas últimas décadas, o mundo tem armazenado uma quantidade incrível de dados, que superam em muito as nossas habilidades para interpretá-los e digeri-los, criando a necessidade de geração de ferramentas e técnicas para automatizar e analisar a base de dados de forma inteligente (FAYYAD, 1996).

Essas técnicas e ferramentas que buscam transformar esses dados armazenados em conhecimento, são o objetivo do campo emergente chamado *Knowledge Discovery in Databases - KDD* (descoberta de conhecimento em bases de dados).

Um crescente número de publicações têm sido dedicadas a este tópico.

Segundo FAYYAD (1996), o termo *Knowledge Discovery in Databases* ou KDD foi criado em 1989 como referência ao processo amplo de encontrar conhecimento em dados e dar ênfase a uma grande aplicação em particular - o método *Data Mining* (Mineração de Dados).

KDD refere-se a todo processo de descoberta de conhecimento útil de dados, enquanto *Data Mining* refere-se a aplicação de algoritmos para extrair modelos dos dados.

Até 1995, muitos autores consideravam os termos KDD e *Data Mining* como sinônimos.

Assim cabe ressaltar que o processo KDD depende de uma nova geração de ferramentas e técnicas de análise de dados e envolve diversas etapas, que serão citadas posteriormente. A principal etapa, núcleo desse processo, chama-se *Data Mining* ou Mineração de Dados, também conhecida como processo de arqueologia de dados, ou reconhecimento de padrões (CHEN; HAN; YU, 1996).

Em 1995, na Conferência Internacional de KDD, em Montreal, foi dada uma definição para cada um dos termos (ADRIAANS e ZANTINGE, 1996):

"...KDD será empregado para todo o processo de extração de conhecimento dos dados. Neste contexto, conhecimento significa relacionamento e padrões entre elementos de dados. O termo Mineração de Dados deveria ser utilizado para os estágios de descoberta do processo de KDD".

A relação existente entre KDD e *Data Mining*, pode ser visualizada através da Figura 3.1.

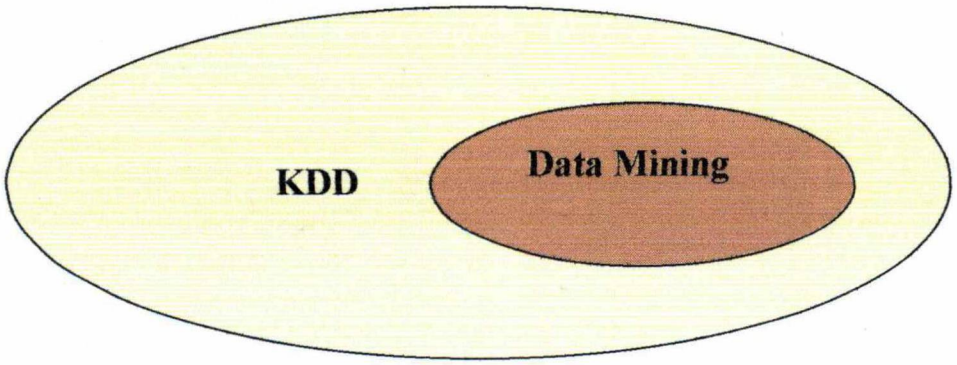


Figura 3.1 - Diferença entre KDD e *Data Mining*

Assim, o processo global para achar e interpretar modelos extraídos de dados é chamado de processo KDD, tipicamente interativo e iterativo, envolvendo repetidas aplicações específicas de métodos ou algoritmos *Data Mining* e a interpretação dos padrões gerados por estes algoritmos (FAYYAD, 1996).

3.2. O PROCESSO DE KDD

KDD é um processo de descoberta de conhecimento em bases de dados que envolvem diversas áreas, tais como: estatística, matemática, banco de dados, inteligência artificial, visualização de dados e reconhecimento de padrões. Este processo utiliza métodos, algoritmos e técnicas oriundos dessas áreas, com o objetivo principal de extrair conhecimento a partir de grandes bases de dados.

O processo de KDD, é um conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados. Segundo FAYYAD (1996), esse conjunto é composto de 5 (cinco) etapas, que são:

- Seleção dos dados;

- Pré-processamento e limpeza dos dados;
- Transformação dos dados;
- ***Data Mining;***
- Interpretação e Avaliação dos resultados.

Essas etapas podem ser visualizadas através da Figura 3.2.

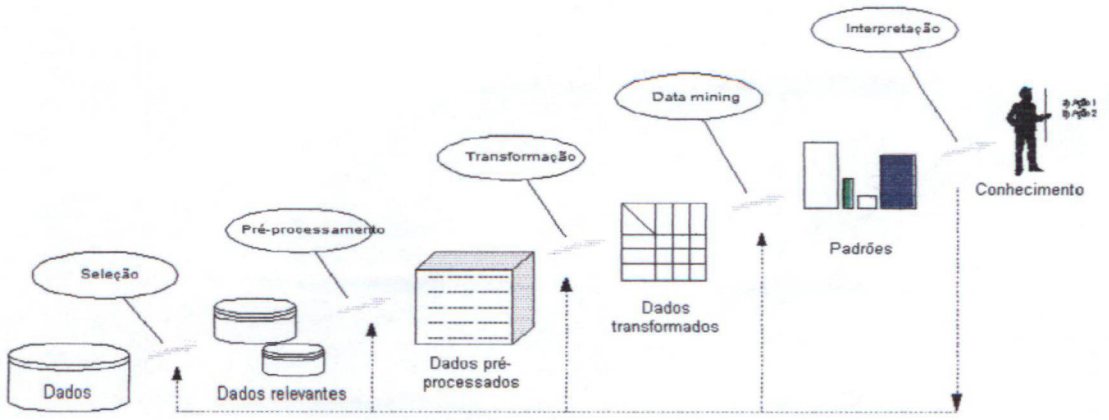


Figura 3.2 - Processo KDD (FAYYAD, 1996)

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 de uma massa de dados, alvo da prospecção. A etapa da limpeza dos dados (*data cleaning)* vem a seguir, através de um pré-processamento dos dados, visando adequá-los aos algoritmos. Isso se faz através da integração de dados heterogêneos, eliminação de incompletude dos dados e outras. Essa etapa pode tomar até 80% do tempo necessário para todo o

processo, devido às bem conhecidas dificuldades de integração de bases de dados heterogêneas (MANNILA, 1996).

Os dados pré-processados devem ainda passar por uma transformação que os armazena adequadamente, visando facilitar o uso das técnicas de *Data Mining.* Nessa fase, o uso de *Data Warehouse* (Armazenamento de Dados) se expande consideravelmente, já que com essa tecnologia as informações estão armazenadas de maneira mais eficiente. Segundo INMON (1997), o *Data Warehouse* é um conjunto de dados, integrado, não volátil e variável em relação ao tempo, dando apoio às decisões gerenciais.

Prosseguindo no processo, chega-se à fase de *Data Mining* especificamente, que começa com a escolha das ferramentas (algoritmos) a serem utilizadas. Essa escolha depende fundamentalmente do objetivo do processo de KDD: classificação, agrupamento, regras associativas, ou outra. De modo geral, na fase de *Data Mining*, ferramentas especializadas procuram padrões nos dados. Essa busca pode ser efetuada automaticamente pelo sistema, de forma livre (*roams* - percorrer/vasculhar o banco de dados) ou interativamente com um analista, responsável pela geração de hipóteses, chamada análise direcionada (*directed analysis*) ou também chamada aprendizado supervisionado (*supervised learning*), onde temos como que um "professor" que "ensina" o sistema indicando, por exemplo, quando uma premissa foi ou não correta.

Diversas ferramentas distintas, como redes neurais, árvores de decisão, sistemas baseados em regras e programas estatísticos, tanto isoladamente quanto em combinação, podem ser então 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 dado aspecto das descobertas, e realimentam o sistema com novos parâmetros. Ao final do processo, o sistema de *Data Mining* gera um relatório das descobertas,

que passa então a ser interpretado pelos analistas de mineração. Somente após a interpretação das informações obtidas encontramos conhecimento.

No passado, a futurologia que envolvia o *Data Mining* sugeria que ele eliminaria a necessidade de analistas estatísticos para construir modelos preditivos. Entretanto, o valor de um analista não pode ser entregue somente por uma ferramenta automática. Analistas serão sempre necessários para avaliar modelos e validar a plausibilidade das predições realizadas. Pelo fato do *software* de *Data Mining* não contar com a experiência e intuição humana para reconhecer a diferença entre uma correlação relevante e irrelevante, analistas estatísticos permanecerão em alta demanda (THEARLING, 2000).

Uma diferença significativa entre *Data Mining* e outras ferramentas de análise está na maneira como exploram as inter-relações entre os dados. As diversas ferramentas de análise disponíveis utilizam um método baseado na verificação, isto é, o usuário constrói hipóteses sobre inter-relações específicas e então verifica ou refuta, através do sistema. Esse modelo torna-se dependente da intuição e habilidade do analista em propor hipóteses interessantes, em manipular a complexidade do espaço de atributos, e em refinar a análise baseado nos resultados de consultas ao banco de dados potencialmente complexas. Já o processo de *Data Mining* fica responsável pela geração de hipóteses, garantindo mais rapidez, acurácia e completude aos resultados.

3.2.1. Seleção dos Dados

Uma vez definido o domínio sobre o qual se pretende executar o processo de descoberta, o próximo passo é selecionar e coletar o conjunto de dados ou variáveis necessárias. A maioria das empresas já possui bases de dados. Porém,

nem sempre todos os dados necessários estão disponíveis em bases adequadas, o que exige um trabalho de compatibilização.

3.2.2. Limpeza dos Dados (*Data Cleaning*)

É a atividade pela qual os ruídos, dados estranhos ou inconsistentes são tratados e onde são estabelecidas as estratégias para resolver os problemas de ausência de dados.

3.2.3. Transformação dos Dados

Nesta fase, como já citado anteriormente, o uso de *Data Warehouse* se expande consideravelmente, já que nessas estruturas as informações estão alocadas da maneira mais eficiente. Um *Data Warehouse* é um repositório de informações para suportar decisões. Ele coleta dados a partir de diversas aplicações de uma organização, integra e organiza os dados em áreas lógicas de assuntos, armazena as informações de forma que elas fiquem acessíveis e compreensíveis a pessoas não técnicas e as disponibiliza da melhor forma possível aos tomadores de decisões, para que possam ser aplicadas técnicas de análise e extração de dados.

3.2.4. *Data Mining*

A atividade de descoberta do conhecimento é uma das mais fascinantes, onde são processados os algoritmos de aprendizado de máquina e de reconhecimento de padrões. A maioria dos métodos de *Data Mining* são baseados

em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação, agrupamento e modelos gráficos.

3.2.5. Interpretação e Avaliação dos Resultados

Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas. Porém, estas formas devem possibilitar uma análise criteriosa para identificar a necessidade de retornar a qualquer um dos estágios anteriores do processo de KDD.

Esta apresentação das atividades pode sugerir que exista uma trajetória linear do processo KDD. No entanto, isso geralmente não se verifica, uma vez que em cada etapa pode ser identificada a necessidade de retorno para cada uma das etapas anteriores. Por exemplo, se na atividade de codificação ou mesmo na de *Data Mining*, é identificado que os dados não estejam plenamente consistentes ou se for verificada a necessidade de um dado que não havia sido previsto anteriormente, isso pode levar ao retorno para a fase de consistência ou mesmo de seleção dos dados.

3.3. ÁREAS RELACIONADAS AO KDD

O processo KDD é interdisciplinar e envolve áreas relativas a aprendizado de máquina, bases de dados, estatística e matemática, sistemas especialistas e visualização de dados.

Este processo utiliza métodos, algoritmos e técnicas oriundos destas diversas áreas, com o objetivo principal de extrair conhecimento a partir de grandes bases de dados.

A relação das áreas no processo KDD pode ser visualizada pela Figura 3.3 a seguir.

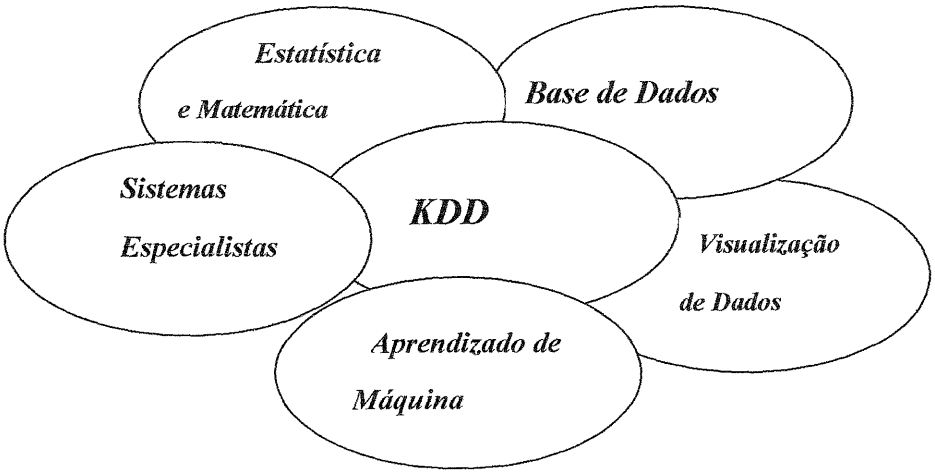


Figura 3.3 - KDD é um campo multi-disciplinar (ADRIAANS e ZANTINGE, 1996)

3.3.1. Aprendizado de Máquina

Nesta área são utilizados modelos cognitivos ou estratégias de aprendizado de máquina, bem como os paradigmas para a aquisição automática de conhecimento.

3.3.2. Bases de Dados

Na área de bases de dados existem tecnologia específicas, bem como uma série de pesquisas que objetivam melhor explorar as características dos dados a serem trabalhados.

3.3.3. Estatística e Matemática

É freqüente que modelos matemáticos ou estatísticos sejam construídos para a geração de regras, padrões e regularidades.

No caso específico da Estatística, essa disponibiliza um grande número de procedimentos técnicos e resultados de testes para as tarefas de *Data Mining,* como, por exemplo, para verificar se estimativas e procedimentos de pesquisa estão consistentes sob determinados critérios de avaliação e identificar o grau de incerteza.

3.3.4. Sistemas Especialistas

Os sistemas especialistas são programas de Inteligência Artificial criados para resolver problemas do mundo real. Inicialmente, estes sistemas ofereciam apenas mecanismos para a representação do conhecimento, raciocínio e explicações. Posteriormente foram incorporadas ferramentas para a aquisição do conhecimento.

3.3.5. Visualização de Dados

A Visualização de Dados assume um papel importante já que em vários momentos existe a necessidade de interação entre o processo de descoberta e o ser humano. Pode-se citar como exemplo, a análise prévia dos dados que vão ou não fazer parte do processo, onde são realizadas algumas consultas usando ferramentas de análise ou mesmo de visualização de dados.

Para a visualização, pode-se recorrer a distintas formas, tais como: gráficos, ícones e figuras.

3.4. *DATA MINING* - DETALHAMENTO

Para que se possa ter uma melhor compreensão a respeito do real contexto sobre *Data Mining,* o mesmo será apresentado nesta seção de forma detalhada.

3.4.1. Introdução ao *Data Mining*

Data mining é a parte mais interessante do processo de KDD, sendo que no contexto negocial é a que mais alavanca e auxilia o empresário a descobrir filões de mercado.

O cérebro humano, comprovadamente, consegue fazer até 8 (oito) comparações ao mesmo tempo. A função do *Data Mining* é justamente ampliar esta comparação para "infinito" e tornar isso visível ao olho humano (POSSAS et all, 1998).

Existe muito conhecimento escondido na imensa quantidade de dados disponíveis nos bancos de dados das empresas. Com o *Data Mining* , pode-se transformar esses dados brutos em informação valiosa para auxiliar o processo decisório.

O *Data Mining* difere de técnicas estatísticas porque, ao invés de verificar padrões hipotéticos, utiliza os próprios dados para descobrir tais padrões.

As bases de dados armazenam conhecimento que podem nos auxiliar a melhorar nossos negócios e as técnicas tradicionais permitem a verificação de hipóteses. Aproximadamente 5% de todas as relações podem ser encontradas por esses métodos. *Data Mining* pode descobrir outras relações anteriormente desconhecidas: os 95% restantes. Em outras palavras, pode-se dizer que técnicas convencionais "falam" à base de dados, enquanto *Data Mining* "ouve" a base de dados. Se você não fizer uma pergunta específica, nunca terá a resposta. *Data Mining* explora as bases de dados através de dezenas de centenas de pontos de vista diferentes. Toda a informação escondida relacionada ao comportamento dos clientes será mapeada e enfatizada (THEARLING, 2000).

Data Mining não substitui técnicas estatísticas tradicionais. Ao invés disto, *Data Mining* é uma extensão dos métodos estatísticos, que são em parte o resultado de uma mudança maior na comunidade estatística. O poder cada vez maior dos computadores com custos mais baixos, aliado à necessidade de análise de enormes conjuntos de dados com milhões de linhas, permitiu o desenvolvimento de técnicas baseadas na exploração de soluções possíveis pela força bruta (THEARLING, 2000).

3.4.2. Algumas Definições de *Data Mining*

São muitas as definições de *Data Mining* encontradas na literatura, sendo que a seguir estão listadas algumas delas.

"*Data Mining* 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 *Data Warehouse* usando técnicas de reconhecimento de padrões, estatísticas e matemáticas" (NIMER & SPANDRI, 1998).

"*Data Mining* é um processo que encontra relações e modelos dentro de um grande volume de dados armazenados em um banco de dados" (RODRIGUES, 2000).

"*Data Mining* é uma técnica para determinar padrões de comportamento, em grandes bases de dados, auxiliando na tomada de decisão" (SILVA, 2000).

"*Data Mining* é um conjunto de técnicas que envolve métodos matemáticos, algoritmos e heurísticas para descobrir padrões e regularidades em grandes conjuntos de dados (POSSAS et all., 1998).

"*Data Mining* é a extração de informações potencialmente úteis e previamente desconhecidas de grandes bancos de dados", relatou a "NEGÓCIOS EXAME", o pesquisador Gregory Piatetsky - Shapiro, uma das maiores autoridades em D*ata Mining* do mundo. "Ele serve para descobrir perfis de consumidores e outros comportamentos que não seriam identificados nem por especialistas" (GUIZZO, 2000).

3.4.3. Objetivos do *Data Mining*

O objetivo principal do *Data Mining* é extrair valiosas informações dos dados, para descobrir o "ouro escondido". Esse "ouro" são as infomações valiosas contidas nos dados. Pequenas mudanças nas estratégias provenientes das descobertas das ferramentas de *Data Mining*, podem traduzir-se em diferenças significativas no caixa da empresa. Com a proliferação dos *Data Warehouses*, as ferramentas de *Data Mining* tornaram-se uma necessidade. Vale a pena lembrar que o uso de um *Data Warehouse* não se faz necessário para a utilização de uma ferramenta de *Data Mining*. Tudo o que é preciso são dados.

Muitas ferramentas tradicionais de análise de dados do tipo geradores de relatórios ou análises estatísticas usam o termo *Data Mining* em seus *softwares* computacionais. Produtos baseados em inteligência artificial também se dizem ferramentas de *Data Mining*. Agora, o que é um verdadeiro *Data Mining* e o que não é? O objetivo principal do *Data Mining* é a descoberta do conhecimento. Sua metodologia extrai informações preditivas das bases de dados.

3.4.4. Origem do *Data Mining*

Data mining é um campo interdisciplinar, que emergiu da interseção entre várias áreas, principalmente aprendizado de máquina (uma subárea da inteligência artificial), estatística e banco de dados, como ilustrado na Figura 3.4 a seguir (FREITAS, 2000).

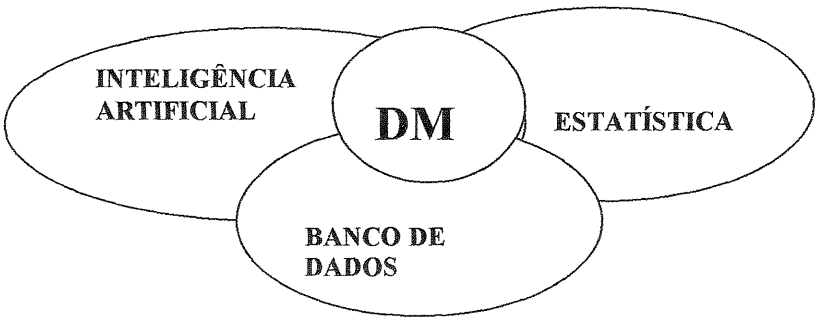


Figura 3.4 - Origem do *Data Mining*

Assim sendo, *Data Mining* é a combinação de diferentes técnicas de sucesso comprovado, como inteligência artificial, estatística e bancos de dados.

3.4.4.1. A Estatística

Sem a estatística não seria possível termos o *Data Mining*, visto que a mesma é a base a partir da qual o *Data Mining* é construído.

A Estatística Clássica envolve conceitos como distribuição normal, variância, análise de regressão, desvio simples, análise de conjuntos, análise de discriminante e intervalos de confiança, todos usados para estudar dados e os relacionamentos entre eles.

Essas são as bases fundamentais onde as mais avançadas análises estatísticas se apóiam. E sem dúvida, na essência das atuais ferramentas e técnicas de *Data Mining*, a análise estatística clássica desempenha um papel fundamental.

3.4.4.2. Inteligência Artificial

Inteligência Artificial ou IA, é uma disciplina construída a partir dos fundamentos da heurística, contrariamente à estatística, tenta imitar a maneira como o homem pensa na resolução dos problemas estatísticos.

Em função dessa abordagem (*approach)*, ela requer um impressionante poder de processamento, que era impraticável até os anos 80, quando os computadores começaram a oferecer um bom poder de processamento a preços mais acessíveis (RODRIGUES, 2000).

O aprendizado de máquina (*machine learning)*, que pode ser melhor descrito como a união entre a estatística e a IA, tenta fazer com que os programas de computador "aprendam" com os dados que eles estudam, tal que esses programas tomem decisões diferentes baseadas nas características dos dados estudados.

3.4.4.3. Banco de Dados

Uma das técnicas mais utilizadas para melhorar a base de dados é o *Data Warehouse,* que é um repositório de dados que é atualizado regularmente sem perder nada - define uma área de interesse e guarda as informações sobre esta área.

Data Warehouse pode também ser definido como um conjunto de tecnologias que permitem converter uma grande quantidade de dados em informação utilizável. Transforma um banco de dados operacional num ambiente que permite o uso estratégico dos dados. É um ambiente e não um produto.

3.4.5. Áreas de Aplicação de *Data Mining*

Técnicas de *Data Mining* têm sido aplicadas com sucesso para a solução de problemas em diversas áreas, como as descritas a seguir:

Vendas

- Retenção de clientes: identificar clientes que podem "migrar" para o competidor 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 (nestes casos, basta enviar mala direta anunciando o produto apenas para aqueles prováveis compradores - mala direta direcionada).

Finanças

- Detectar padrões de fraudes no uso dos cartões de crédito;
- Identificar os consumidores que estão tendendo a mudar a companhia do cartão de crédito;
- Identificar regras de estocagem a partir dos dados do mercado;
- Encontrar correlações escondidas nas bases de dados.

Seguros e Planos de Saúde

- Determinar quais procedimentos médicos são requisitados ao mesmo tempo;
- Prever quais consumidores comprarão novas apólices;

- Identificar comportamentos fraudulentos.

Transporte

- Determinar a distribuição dos horários entre os vários caminhos;
- Analisar padrões de sobrecarga.

Medicina

- Caracterizar o comportamento dos pacientes para prever novas consultas;
- Identificar terapias de sucessos 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 (particularmente em celulares), dentre um enorme número de ligações efetuadas pelos clientes.

Mercado Financeiro

- Prever quais ações irão subir ou descer na bolsa de valores, em função de dados históricos com preços de ações e valores de índices financeiros.

Segundo o engenheiro Alessandro Zanasi, formado em Engenharia Nuclear, atualmente consultor da IBM no *Bologna Data Mining Center,* em entrevista concedida ao Planeta COPPE (2001), na Itália, uma rede de supermercados, descobriu que os clientes que compravam fraldas também compravam muita bebida alcoólica. A princípio não havia nenhuma razão lógica para tal. No entanto, utilizando o *Data Mining* descobriram que os compradores eram casais jovens que não podiam sair de casa com tanta freqüência por causa dos filhos pequenos.

Passavam a receber freqüentemente, intensificando a vida social em casa. Sabendo disso a rede de supermercados trabalhou uma série de ofertas direcionadas especificamente a este consumidor

Na mesma entrevista, o engenheiro Alessandro Zanasi relatou um caso bastante curioso. Dois times de basquete ligados à Associação Nacional de Basquete dos EUA (NBA), o *New York Knicks* e o *Miami Heat* solicitaram uma consultoria à IBM, nos Estados Unidos, para descobrir as estratégias dos times adversários. Os treinadores só conceberam o esquema tático e colocaram o time em campo depois que o *Data Mining* revelou os pontos fracos dos adversários.

3.4.6. Características Desejáveis do Conhecimento a ser Descoberto por *Data Mining*

Segundo FREITAS (2000), idealmente, o conhecimento a ser descoberto deve satisfazer três propriedades, a saber:

- correto (tanto quanto possível);
- compreensível por usuários humanos;
- interessante / útil / novo (surpreendente).

3.4.7. Características Desejáveis do Método de Descoberta de Conhecimento por *Data Mining*

Ainda, segundo FREITAS (2000), o método de descoberta do conhecimento deve apresentar as seguintes características:

- eficiente (rápido);
- genérico (aplicável a vários tipos de dados);
- flexível (facilmente modificável).

3.5. PRINCIPAIS TÉCNICAS DE *DATA MINING*

O *Data Mining* é um campo que compreende atualmente muitas ramificações importantes. Cada tipo de tecnologia tem suas próprias vantagens e desvantagens, do mesmo modo que nenhuma ferramenta consegue atender todas as necessidades em todas as aplicações.

Dentre as técnicas de *Data Mining*, destacam-se as apresentadas nas seções a seguir.

3.5.1. Árvores de Decisão

Árvore de Decisão é um método adequado quando o objetivo do *Data Mining* é classificação de dados ou predição de saídas. É conveniente usar árvore de decisão quando o objetivo for categorizar dados de arquivos. Também é uma boa escolha quando o objetivo é gerar regras que podem ser facilmente entendidas, explicadas e traduzidas para linguagem natural.

3.5.2. Redes Neurais

As Redes Neurais tentam construir representações internas de modelos ou padrões detectados nos dados, mas essas representações não são apresentadas para o usuário.

Estruturalmente, uma Rede Neural consiste em um número de elementos interconectados (chamados neurônios) organizados em camadas que aprendem pela modificação da conexão que conectam as camadas.

As Redes Neurais Artificiais utilizam um conjunto de elementos de processamento (ou nós) análogos aos neurônios no cérebro. Estes elementos de processamento são interconectados em uma rede que pode identificar padrões nos dados uma vez expostos aos mesmos, ou seja, a rede aprende através da experiência, tais como as pessoas (DIN, 1998).

3.5.3. Análise de Agrupamento

Esta técnica agrupa informações homogêneas de grupos heterogêneos entre os demais e aponta o item que melhor representa cada grupo, permitindo desta forma que se consiga perceber a característica de cada grupo. Desse modo, objetos dentro do mesmo grupo são os mais semelhantes possíveis, enquanto que objetos de grupos diferentes são os mais diferentes possíveis.

Por exemplo, suponha que os objetos sejam clientes, e que se tenha vários atributos descrevendo cada cliente, tais como a idade, faixa de salário, sexo e outros. Analisando esses dados, um sistema de *Data Mining* pode, por exemplo, criar um grupo de clientes com idade baixa e faixa de salário baixa, outro grupo de clientes com idade alta e faixa de salário alto e assim por diante. Essa diferenciação dos clientes em grupos pode ser bastante útil, já que clientes de grupos diferentes, presumidamente, tendem a ter comportamentos de compra bem diferentes (FREITAS, 2000).

3.5.4. Indução de Regras

A Indução de Regras (*Rule Induction)* se refere à detecção de tendências dentro de grupos de dados ou de "regras" sobre os dados. As regras são, então, apresentadas aos usuários como uma lista "não encomendada", ou seja, sem que obedecam algum critério previamente estabelecido.

Indução de Regras é o processo de analisar uma série de dados e, a partir dela, gerar padrões. O processo é, em sua essência, semelhante àquilo que um analista humano faria em uma análise exploratória.

Consiste na descoberta de regras de previsão, do tipo SE...ENTÃO, onde a parte SE (a "condição") da regra especifica alguns valores de atributos previsores e a parte ENTÃO da regra prevê um valor para um determinado atributo cuja previsão é desejada. Por exemplo, suponha que se tenha um banco de dados de vendas de produtos, com dados sobre produtos vendidos e os clientes que compraram aqueles produtos. Assuma que os dados incluem atributos tais como a idade e sexo do cliente e o tipo do produto comprado. Analisando esses dados, um sistema de *Data Mining* poderia descobrir uma regra de previsão do tipo SE ... ENTÃO, tal como: SE (idade_cliente < 18) E (sexo_cliente = "M") ENTÃO (produto_comprado_videogame) (FREITAS, 2000).

Idealmente as regras descobertas deveriam satisfazer três propriedades, a saber:

(a) fazerem previsões corretas, ou seja, na maioria das vezes que a parte "SE" da regra é verdadeira, a parte "ENTÃO" da regra também é verdadeira;

(b) serem compreensíveis para o usuário, ou seja, as regras representam conhecimento em um alto nível de abstração, tal como a regra acima, ao invés de equações matemáticas complexas e não compreensíveis pelo usuário;

(c) serem úteis para a tomada de decisão, o que está relacionado ao fato da regra expressar conhecimento novo ou surpreendente para o usuário. No exemplo acima, o usuário poderia usar a regra descoberta para, por exemplo, fazer uma mala direta direcionada, enviando uma propaganda de um novo videogame apenas para clientes que têm menos de 18 anos e são do sexo masculino (FREITAS, 2000).

Vários algoritmos e índices são usados para executar esse processo. Na Indução de Regras, a grande maioria do processo é feito pela máquina e uma pequena parte é feita pelo usuário.

3.5.5. Análise Estatística de Séries Temporais

A estatística é a mais antiga tecnologia em *Data Mining*, e é parte da fundamentação básica de todas as outras tecnologias. Ela incorpora um envolvimento muito forte do usuário, exigindo engenheiros experientes, para construir modelos que descrevam o comportamento dos dados através dos métodos clássicos de matemática. Interpretar os resultados dos modelos requer especialistas (*expertises).* O uso de técnicas estatísticas também requer um trabalho muito forte de máquinas/engenheiros.

A análise de séries temporais é um exemplo disso, apesar de freqüentemente ser confundida como um gênero mais simples de *Data Mining* chamado previsão (*Forecasting).*

Enquanto que a análise de séries temporais é um ramo altamente especializado da estatística, o *Forecasting* é, de fato, uma disciplina muito menos rigorosa, que pode ser satisfeita, embora com menos segurança, através da maioria das outras técnicas de *Data Mining.*

3.5.6. Visualização

As técnicas de Visualização são um pouco mais difíceis de definir, sendo que muitas pessoas a definem como "ferramentas complexas de visualização", enquanto outras como simplesmente a capacidade de geração de gráficos.

Nos dois casos, a Visualização mapeia o dado que está sendo minerado de acordo com dimensões especificadas. Nenhuma análise é executada pelo programa de *Data Mining* além da manipulação da estatística básica. O usuário, então, interpreta o dado através do monitor de vídeo.

3.6. ETAPAS DO *DATA MINING*

A implementação de um sistema de *Data Mining* pode ser dividida em seis fases interdependentes para que o mesmo atinja seus objetivos finais, descritas a seguir.

3.6.1. Entendimento do Problema

A fase inicial do projeto deve ter por objetivo identificar as metas e necessidades a partir de uma perspectiva do problema, e então convertê-las para uma aplicação de *Data Mining* e um plano inicial de "ataque" ao problema.

3.6.2. Entendimento dos dados

Esta fase tem como atividade principal extrair uma amostra dos dados a serem usados e avaliar o ambiente em que os mesmos se encontram.

3.6.3. Preparação dos dados

Criação de programas de extração, limpeza e transformação dos dados para uso pelos algoritmos de *Data Mining*. É nessa etapa que os dados são adaptados para serem então inseridos no algoritmo escolhido para processamento.

3.6.4. Modelagem do Problema

Seleção do(s) algoritmo(s) dentre os apresentados a serem utilizados e efetivo processamento do modelo. Alguns algoritmos necessitam dos dados em formatos específicos, o que acaba causando vários retornos à fase de preparação dos dados.

3.6.5. Avaliação do Modelo

Ao final da fase de modelagem, vários modelos devem ter sido avaliados sob a perspectiva do analista responsável. Agora, o objetivo passa a ser avaliar os modelos com a visão do problema, certificando-se que não existem falhas ou contradições com relação às regras do problema.

3.6.6. Divulgação ou Publicação do Modelo

A criação e validação do modelo permite avançar mais um passo, no sentido de tornar a informação gerada acessível. Isto pode ser feito de várias maneiras, desde a criação de um *software* específico para tal, até a publicação de um relatório para uso interno.

3.7. VANTAGENS DO *DATA MINING*

O uso de *Data Mining* para construção de um modelo traz as seguintes vantagens:

- **Modelos são de fácil compreensão:** pessoas sem conhecimento estatístico (por exemplo, analistas financeiros ou pessoas que trabalham com *data base marketing*) podem interpretar o modelo e compará-lo com suas próprias idéias. O usuário ganha mais conhecimento sobre o comportamento do cliente e pode usar esta informação para otimizar os processos dos negócios.

- **Grandes bases de dados podem ser analisadas:** grandes conjunto de dados, de até vários gigabytes de informação podem ser analisados com *Data Mining*.

- ***Data Mining* descobre informações não esperadas:** como muitos modelos diferentes são validados, alguns resultados inesperados podem surgir. Em diversos estudos, descobriu-se que combinações de fatores particulares apresentaram resultados inesperados.

- **Variáveis não necessitam de recodificação:** *Data Mining* lida tanto com variáveis numéricas (quantitativas) quanto categóricas (qualitativas). Estas

variáveis aparecem no modelo exatamente da mesma forma em que aparecem na base de dados.

- **Modelos são precisos:** os modelos obtidos por *Data Mining* são validados por técnicas de estatística. Desta forma, as predições feitas por modelos são precisas.

CAPÍTULO IV

4. DESCRIÇÃO DAS TÉCNICAS APLICADAS AO PROBLEMA

4.1. INTRODUÇÃO

A finalidade deste capítulo é mostrar com um maior nível de detalhamento as técnicas utilizadas neste trabalho com o objetivo de se atingir as metas propostas, já descritas no Capítulo I, no item 1.1.

4.2. REDES NEURAIS

O desenvolvimento das Redes Neurais Artificiais começaram há aproximadamente 60 anos, motivado por um desejo de tentar compreender o cérebro e emular algumas de suas forças (FAUSETT, 1995).

4.2.1. Histórico

A história registra que as primeiras informações sobre a neuro computação datam de 1943, em artigos de McCulloch (neurobiologista Warren McCulloch de *Massachusets Institute of Tecnology - MIT*) e Pitts (matemático Walter Pitts da Universidade de *Illinois*), os quais, sugeriam a construção de uma máquina baseada ou inspirada no cérebro humano. Eles fizeram uma analogia entre células nervosas biológicas e um processo eletrônico num trabalho publicado sobre

"neurônios formais". O trabalho mostrava que uma coleção de neurônios era capaz de calcular certas funções lógicas.

Muitos outros artigos e livros surgiram desde então, porém, por um longo período de tempo, poucos resultados foram obtidos, até que em 1949, Donald Hebb escreveu um livro intitulado "*The Organization of Behavior*" ("A Organização do Comportamento") que perseguia a idéia de que o condicionamento psicológico clássico está presente em qualquer animal pelo fato de que esta é uma propriedade de neurônios individuais. De acordo com sua teoria, "se um neurônio A é repetidamente estimulado por outro neurônio B, ao mesmo tempo que ele está ativo, ele ficará mais sensível aos estímulos de B, e a conexão sináptica de B para A será mais eficiente. Deste modo, B achará mais fácil estimular A para produzir uma saída". Suas idéias não eram completamente novas, mas Hebb foi o primeiro a propor uma lei de aprendizagem específica para as sinapses dos neurônios. Este primeiro e corajoso passo serviu de inspiração para que muitos outros pesquisadores perseguissem a mesma idéia (TATIBANA, 2000).

Em 1951 foi construído o primeiro neuro computador, denominado *Snark*, por Mavin Minsky. O *Snark* operava com sucesso a partir de um ponto inicial, ajustando seus pesos automaticamente, entretanto, ele nunca executou qualquer função de processamento de informação interessante, mas serviu de inspiração para as idéias de estruturas que o sucederam (TATIBANA, 2000).

Ainda, segundo TATIBANA (2000), em 1956 no *Darthmouth College* nasceram os dois paradigmas da Inteligência Artificial, a simbólica e a conexionista. A Inteligência Artificial Simbólica tenta simular o comportamento inteligente humano desconsiderando os mecanismos responsáveis por tal. Já a Inteligência Artificial Conexionista acredita que construindo-se um sistema que simule a estrutura do cérebro, este sistema apresentará inteligência, ou seja, será capaz de aprender, assimilar, errar e aprender com seus erros.

O primeiro neuro computador a obter sucesso *(Mark I Perceptron)* surgiu em 1958, criado por Frank Rosenblatt. Devido a profundidade de seus estudos, suas contribuições técnicas e de sua maneira moderna de pensar, muitos o vêem como o fundador da neuro computação na forma em que a temos hoje. Seu interesse inicial para a criação do Perceptron era o reconhecimento de padrões.

Após Rosenblatt, Bernard Widrow, com a ajuda de alguns estudantes, desenvolveram um novo tipo de elemento de processamento de redes neurais, chamado de *ADALINE (ADAptative Linear NEtwork)*, o qual dispunha de uma poderosa estratégia de aprendizado, que diferentemente do Perceptron, ainda permanece em uso. Widrow também fundou a primeira companhia de *hardware* de neurocomputadores e componentes.

Infelizmente, os anos seguintes foram marcados por um entusiasmo exagerado de muitos pesquisadores, que passaram a publicar mais e mais artigos e livros que faziam uma previsão pouco confiável para a época, sobre máquinas tão poderosas quanto o cérebro humano que surgiriam em um curto espaço de tempo. Isto tirou quase toda a credibilidade dos estudos desta área e causou grandes aborrecimentos aos técnicos de outras áreas.

Um período de poucas pesquisas seguiu-se durante 1967 a 1982, devido aos fatos ocorridos anteriormente, porém com destaque para pesquisas realizadas por MinsKy & Papert, em 1969. Entretanto, aqueles que pesquisavam nesta época, e todos os que se seguiram no decorrer de treze anos conseguiram novamente estabelecer um campo concreto para o renascimento da área.

Nos anos 80, muitos dos pesquisadores foram bastante corajosos e passaram a publicar diversas propostas para a exploração do desenvolvimento de redes neurais bem como suas aplicações. Porém, talvez o fato mais importante

deste período tenha ocorrido quando Ira Skurnick, um administrador de programas da *DARPA (Defense Advanced Research Projects Agency)* decidiu ouvir os argumentos da neuro computação e seus projetistas, e divergindo dos caminhos tradicionais dos conhecimentos convencionais, fundou em 1983 pesquisas em neuro computação. Este ato não só abriu as portas para a neuro computação, como também deu a *DARPA* o status de uma das líderes mundiais em se tratando de tendência tecnológica.

Outra "potência" que emergiu neste período foi John Hopfield, renomado físico de reputação mundial, se interessou pela neuro computação, e escreveu artigos que percorreram o mundo todo, persuadindo centenas de cientistas, matemáticos e tecnólogos altamente qualificados a se unirem a esta nova área emergente.

Em 1986, Rumelhart, Hinton e Williams introduziram o poderoso método *Backpropagation.* Neste mesmo ano, este campo de pequisa "explodiu" com a publicação do livro "*Parallel Distributed Procesing*" ("Processamento Distribuído Paralelo") editado por David Rumelhart e James McClelland (MENDES FILHO, 1997).

Em 1987, ocorreu em São Francisco a primeira conferência de redes neurais em tempos modernos, a *IEEE International Conference on Neural Networks*, e também foi formada a *International Neural Networks Society (INNS).* A partir destes acontecimentos ocorreram a fundação do *INNS journal* em 1989, seguido do *Neural Computation* e do *IEEE Transactions on Neural Netwoerks* em 1990.

Desde 1987, muitas universidades anunciaram a formação de institutos de pesquisa e programas de educação em neuro computação.

Entretanto, para se ter um histórico completo, devem ser citados alguns pesquisadores que realizaram, nos anos 60 e 70, importantes trabalhos sobre modelos de redes neurais em visão, memória, controle e auto-organização, como: Amari, Anderson, Cooper, Cowan, Fukushima, Grossberg, Kohonen, von der Malsburg, Werbos e Widrow (MENDES FILHO, 1997).

4.2.2. Resumo dos Fatos Históricos em Ordem Cronológica

A seguir são apresentados um resumo dos fatos históricos em ordem cronológica:

- **1943** - McCULLOUCH e PITTS estabeleceram as bases da neurocomputação, com modelos matemáticos. O trabalho fazia uma analogia entre células vivas e o processo eletrônico, simulando o comportamento do neurônio natural, onde o neurônio possuia apenas uma saída, que era uma função da entrada *threshold* e da soma do valor de suas diversas entradas. A idéia de funcionamento desse neurônio artificial pode ser visualizada pela Figura 4.1. a seguir.

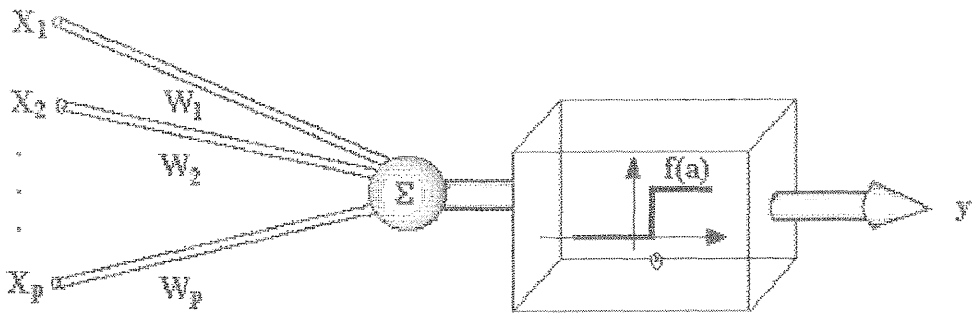


Figura 4.1 - Neurônio artificial projetado por McCulloch & Pitts

- **1949** - HEBB traduziu matematicamente a sinapse dos neurônios biológicos. O psicólogo Donald Hebb demonstrou que a capacidade de aprendizagem em redes neurais vem da alteração da eficiência sináptica, isto é, a conexão somente é reforçada se tanto as células pré-sinápticas quanto as pós-sinápticas estiverem excitadas.

- **1951** - MINSKI constrói o *Snark*, primeiro neurocomputador com capacidade de aprendizado, ou seja, ajustava automaticamente os pesos entre as sinapses, porém não executou nenhuma função útil.

- **1957** - ROSENBLATT concebeu o "perceptron", que era uma rede neural de duas camadas, usado no reconhecimento de caracteres.

- **1958** - ROSENBLATT mostrou em seu livro (*Principles of Neurodynamics*) o modelo dos "perceptrons". Nele, os neurônios eram organizados em camada de entrada e saída, onde os pesos das conexões eram adaptados a fim de se atingir a eficiência sináptica, conforme esquematizado na Figura 4.2.

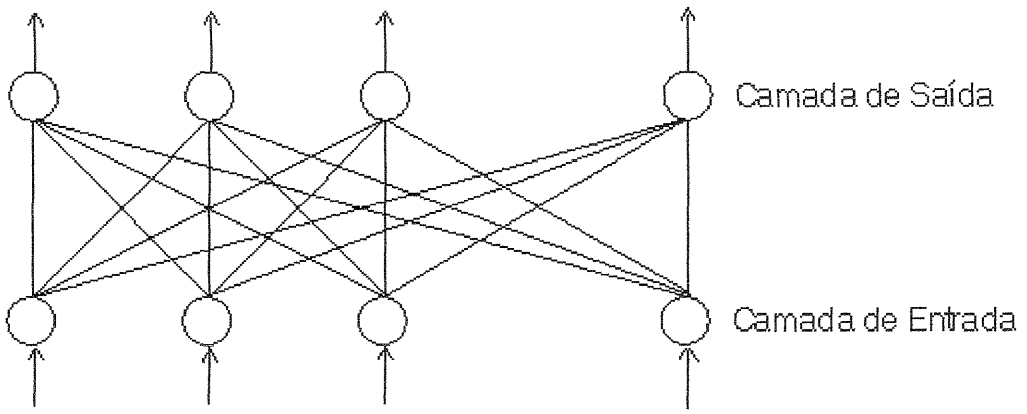


Figura 4.2 - Rede de perceptrons proposta por Rosenblatt

- **1960** - WIDROW e HOFF propuseram a rede *ADALINE (ADAptative LInear NEtwork)* e o *MADALINE (MAny ADALINE)* perceptron. O *ADALINE/MADALINE* utilizou saídas analógicas em uma arquitetura de três camadas, conforme apresentado na Figura 4.3.

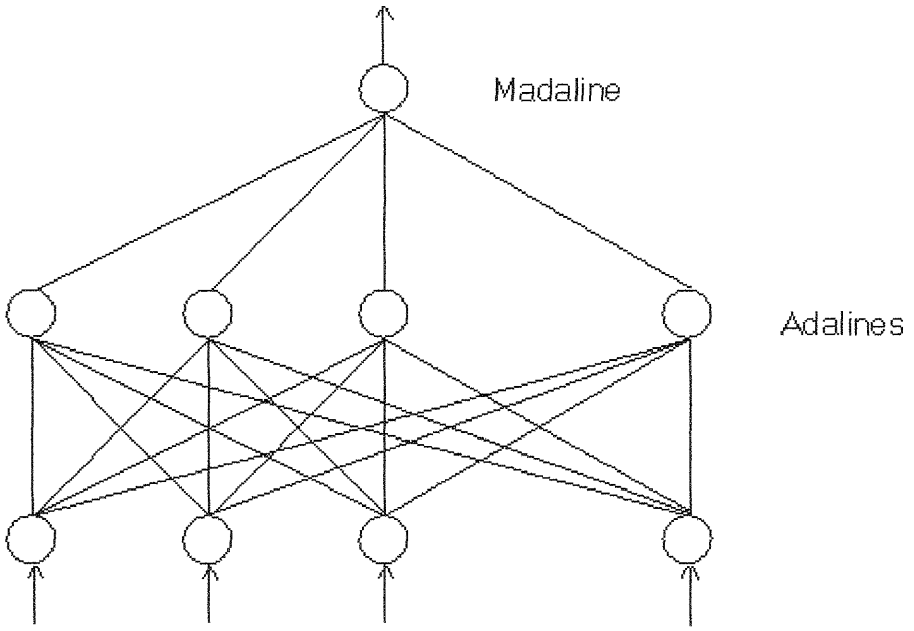


Figura 4.3 - Redes *ADALINE* e *MADALINE*

- **1962** - WIDROW fundou a primeira empresa de circuitos neurais digitais, a *Memistor Corporation.*

- **1967** - Ocorreu a finalização da concessão de verbas destinadas à pesquisa de redes neurais (TATIBANA, 2000)

- **1974** - WERBOS lançou as bases para o algoritmo de retropropagação (*Back-Propagation*).

- **1986** - RUMELHART, HINTON e WILLIAMS introduziram o poderoso método *Back-Propagation,* sintetizado na Figura 4.4.

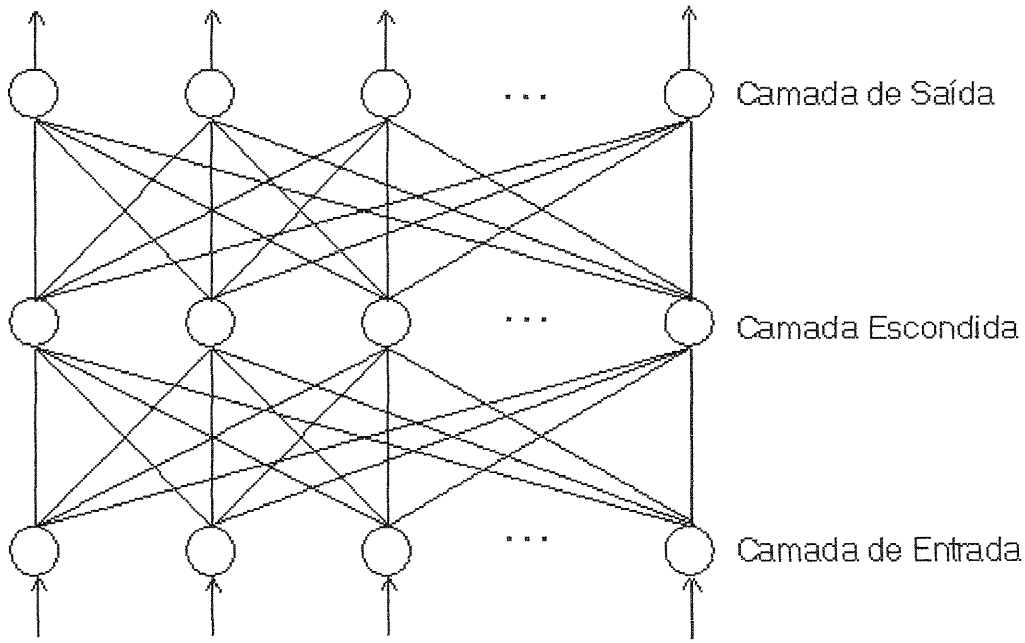


Figura 4.4 - Estrutura do Método *Back-Propagation*

4.3. O NEURÔNIO BIOLÓGICO

A célula nervosa, ou, simplesmente neurônio, é o principal componente do sistema nervoso. Considerada sua unidade anatomo-fisiológica, estima-se que no

cérebro humano existam aproximadamente 10 bilhões destas células, responsáveis por todas as funções do sistema.

Existem diversos tipos de neurônios, com diferentes funções dependendo da sua localização e estrutura morfológica, mas que, em geral, constituem-se dos mesmos componentes básicos:

- o corpo do neurônio (soma) constituído de núcleo e pericário, que dá suporte metabólico à toda célula;

- o axônio (fibra nervosa) prolongamento único e grande que aparece no soma. É responsável pela condução do impulso nervoso para o próximo neurônio, podendo ser revestido ou não por mielina (bainha axonial), célula glial especializada;

- os dendritos que são prolongamentos menores em forma de ramificações (arborizações terminais) que emergem do pericário e do final do axônio, sendo, na maioria das vezes, responsáveis pela comunicação entre os neurônios através das sinapses. Basicamente, os dentritos tem por função, receber os estímulos transmitidos pelos outros neurônios.

A sinapse é a estrutura dos neurônios através da qual ocorrem os processos de comunicação entre os mesmos, ou seja, onde ocorre a passagem do sinal neural (transmissão sináptica) através de processos eletroquímicos específicos, isso graças a certas características particulares da sua constituição. Em outras palavras, a sinapse é a região onde dois neurônios entram em contato e através da qual os impulsos nervosos são transmitidos entre eles.

Em uma sinapse os neurônios não se tocam, permanecendo um espaço entre eles denominado fenda sináptica, onde um neurônio pré-sináptico liga-se a

um outro denominado neurônio pós-sináptico. O sinal nervoso (impulso), que vem através do axônio da célula pré-sináptica chega em sua extremidade e provoca na fenda a liberação de neurotransmissores depositados em bolsas chamadas de vesículas sinápticas. Este elemento químico se liga quimicamente a receptores específicos no neurônio pós-sináptico, dando continuidade à propagação do sinal.

Um neurônio pode receber ou enviar entre 1.000 a 100.000 conexões sinápticas em relação a outros neurônios, dependendo de seu tipo e localização no sistema nervoso. O número e a qualidade das sinapses em um neurônio pode variar, entre outros fatores, pela experiência e aprendizagem, demonstrando a capacidade plástica do sistema nervoso.

4.3.1. Como Funciona o Sistema Nervoso Biológico

O sistema nervoso detecta estímulos externos e internos, tanto físicos quanto químicos, e desencadeia as respostas musculares e glandulares. Assim, é responsável pela integração do organismo com o seu meio ambiente.

Ele é formado, basicamente, por células nervosas, que se interconectam de forma específica e precisa, formando os chamados circuitos neurais. Através desses circuitos, o organismo é capaz de produzir respostas estereotipadas que constituem os comportamentos fixos e invariantes (por exemplo, os reflexos), ou então, produzir comportamentos variáveis em maior ou menor grau.

Todo ser vivo dotado de um sistema nervoso é capaz de modificar o seu comportamento em função de experiências passadas. Essa modificação comportamental é chamada de aprendizado, e ocorre no sistema nervoso através da propriedade chamada plasticidade cerebral (TAFNER, 1998).

4.4. O NEURÔNIO ARTIFICIAL

O neurônio artificial é uma estrutura lógico-matemática que procura simular a forma, o comportamento e as funções de um neurônio biológico. Assim sendo, os dendritos foram substituídos por *entradas*, cujas ligações com o corpo celular artificial são realizadas através de elementos chamados de *peso* (simulando as sinapses). Os estímulos captados pelas entradas são processados pela *função de soma*, e o limiar de disparo do neurônio biológico foi substituído pela *função de transferência* (TAFNER, 1998).

A Figura 4.5. a seguir mostra um neurônio biológico e a Figura 4.6. apresenta os componentes do neurônio artificial.

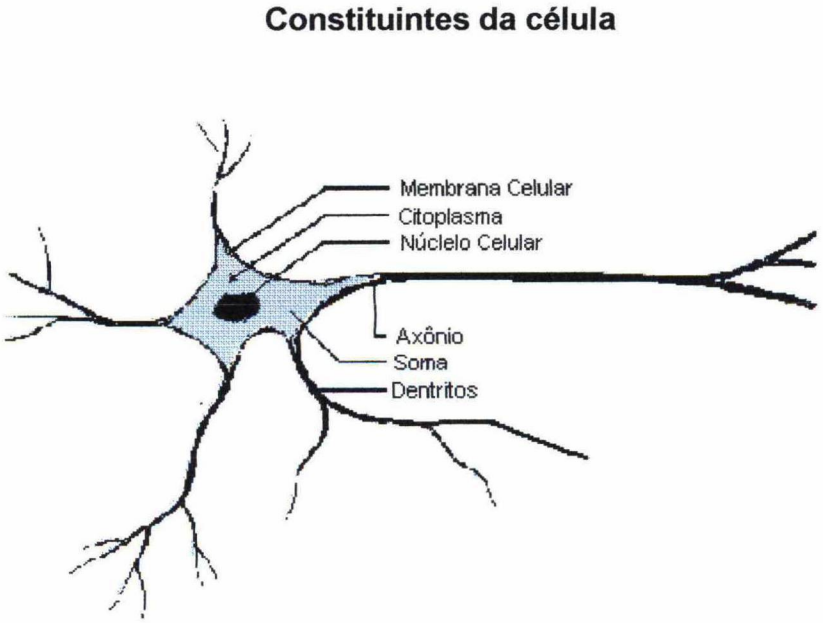


Figura 4.5. - NEURÔNIO BIOLÓGICO

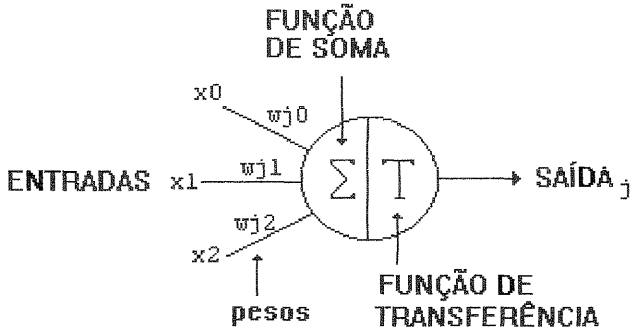


Figura 4.6. - NEURÔNIO ARTIFICIAL

A função básica de um neurônio através de treinamento supervisionado é, depois de acumular o valor somado dos produtos ocorridos entre as entradas e os pesos, processar esse valor através de uma função de ativação e passá-lo adiante através da saída (esse processo é chamado de *função de transferência)*.

Combinando diversos neurônios artificiais forma-se o que é chamado de Rede Neural Artificial.

4.5. REDES NEURAIS ARTIFICIAIS

As redes neurais artificiais consistem em um método para solucionar problemas da área de inteligência artificial, através da construção de um sistema que tenha circuitos que simulem o cérebro humano, inclusive seu comportamento, ou seja, aprendendo, errando e fazendo descobertas. São mais que isso, são técnicas computacionais que apresentam um modelo inspirado na estrutura neural dos organismos inteligentes, que adquirem conhecimento através da experiência.

Uma Rede Neural Artificial pode ter centenas ou até milhares de unidades de processamento, enquanto que o cérebro de um mamífero pode ter muitos bilhões de neurônios.

As redes neurais utilizam um conjunto de elementos de processamento (ou nós) análogos aos neurônios no cérebro. Estes elementos de processamento em uma rede neural são interconectados podendo identificar padrões nos dados apresentados, uma vez expostos aos mesmos, pois as redes aprendem através da experiência, tais como as pessoas. Esta característica distingue redes neurais de tradicionais programas computacionais, que simplesmente seguem instruções em uma ordem seqüencial fixa (DIN, 1998).

Assim como o sistema nervoso é composto de bilhões de células nervosas, a rede neural artificial também seria formada por unidades que nada mais são do que pequenos módulos (ou unidades de processamento ou nós) que simulam o funcionamento de um neurônio. Estes módulos devem funcionar de acordo com os elementos em que foram inspirados, recebendo e retransmitindo informações.

Numa Rede Neural Artificial as entradas, simulando uma área de captação de estímulos, podem ser conectadas em muitos neurônios, resultando, assim, em uma série de saídas, onde cada neurônio representa uma saída. Essas conexões, em comparação com o sistema biológico, representam o contato dos dentritos com outros neurônios, formando assim as sinapses. A função da conexão em si é tornar o sinal de saída de um neurônio em um sinal de entrada de outro, ou ainda, orientar o sinal de saída para o mundo externo (mundo real). As diferentes possibilidades de conexões entre as camadas de neurônios podem ter, em geral, *n* números de estruturas diferentes (TAFNER, 1998). Um exemplo é mostrado pela Figura 4.7.

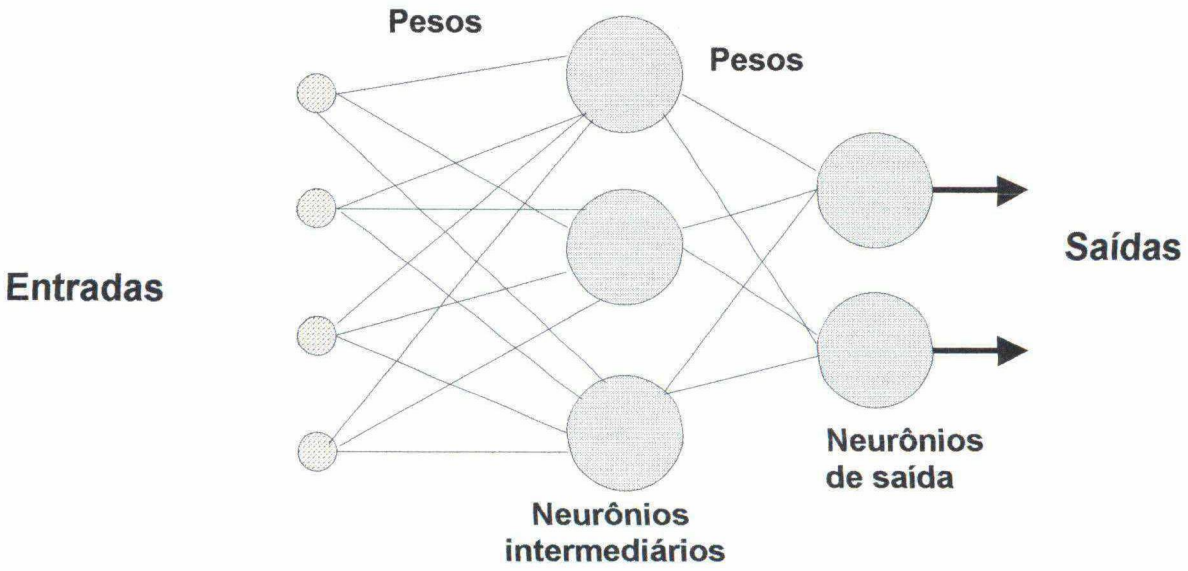


Figura 4.7. - Exemplo de uma Rede Neural Artificial

Usualmente, trabalha-se com três camadas, que são classificadas em:

- Camada de Entrada: onde os padrões são apresentados à rede;
- Camadas Intermediárias ou Ocultas: onde é feita a maior parte do processamento, através das conexões ponderadas; podem ser consideradas como extratoras de características;
- Camada de Saída: onde o resultado final é concluído e apresentado.

As variantes de uma rede neural são muitas, e podem ser "trabalhadas" conforme a aplicação. O que faz as redes neurais diferirem entre si, são os tipos de conexões e formas de treinamento. Basicamente, os itens que compõem uma rede neural e, portanto, sujeitos a modificações, são os seguintes:

- forma de conexões entre camadas;
- número de camadas intermediárias;
- quantidade de neurônios em cada camada;
- função de transferência;
- algoritmo de aprendizado.

4.5.1. Características Gerais das Redes Neurais

Uma Rede Neural Artificial é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente são conectadas por canais de comunicação que estão associados a determinados pesos. Os pesos do neurônio artificial nada mais são do que um modelo para simular os dentritos, que são os responsáveis pelas sinapses. São os pesos que alterando os seus valores representativos durante os estímulos, influenciam o resultado do sinal de saída (TAFNER, 1998).

As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento "inteligente" de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede.

O funcionamento de uma Rede Neural Artificial, é demonstrado através da Figura 4.8.

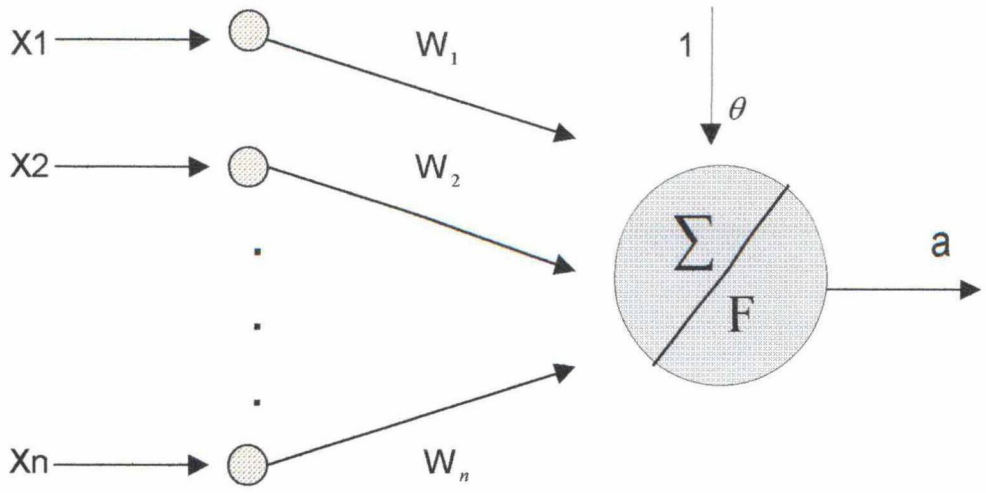


Figura 4.8. - Funcionamento de um neurônio artificial

Segundo KRÖSE e SMAGT (1993), o vetor $\underline{x}$ que representa um conjunto de "n" entradas, é multiplicado por um vetor de pesos, $\underline{w}$, e o produto, $p = \underline{x}\,\underline{w}$, é aplicado aos canais de entrada do neurônio. A soma de todas as entradas ponderadas é então processada por uma função de ativação, $F(\underline{x})$, que vai produzir o sinal de saída "**a**", do neurônio:

$$\mathbf{a} = \mathbf{F}\left(\sum_{i=0}^{n-1} x_i w_i + \theta\right)$$

O parâmetro θ é um valor *threshold* adicionado a soma ponderada, e em alguns casos é omitido, enquanto que em outros é considerado como o valor peso cujo correspondente valor de entrada é sempre igual a 1.

Segundo STEINER (1999), o papel de θ, chamado de *bias* ou vício, é aumentar o número de graus de liberdade disponíveis no modelo, permitindo que a rede neural tenha maior capacidade de se ajustar ao conhecimento a ela fornecido.

4.5.2. A Função de Ativação de uma Rede Neural

A função de ativação é muito importante para o comportamento de uma Rede Neural porque é ela que define a saída do neurônio artificial e portanto o caminho pelo qual a informação é conduzida (STEINER, 1999).

É através de uma função de ativação que são calculadas as respostas geradas pelas unidades. Existem vários tipos de funções de ativação, sendo que as mais comuns são as descritas a seguir (STEINER, 1999).

- FUNÇÃO PASSO, que produz uma saída binária, e embora seja similar aos neurônios reais, é inadequada para o algoritmo de aprendizagem;

- FUNÇÃO LINEAR, que elimina a descontinuidade em $x = \theta$;

- FUNÇÃO SIGMOIDAL, que adiciona alguma não-linearidade (STEINER, 1999).

4.5.3. Classificação das Redes Neurais Artificiais

Existem diversos tipos de Redes Neurais Artificiais e diferentes maneiras de classificá-las. Talvez a mais importante seja quanto à forma de aprendizado ou treinamento, que pode ser supervisionado ou não-supervisionado.

No aprendizado supervisionado são apresentados à rede padrões de entrada e suas saídas (respostas desejadas). Durante este processo, a rede realiza um ajustamento dos pesos das conexões entre os elementos de processamento, segundo uma determinada regra de aprendizagem, até que o erro calculado em função das saídas geradas pela rede alcancem um valor mínimo desejado (SIMEÃO, 1999). Dentre os diversos tipos de rede que apresentam aprendizagem supervisionada, pode-se destacar por exemplo, *perceptron, adaline* e *madaline.*

O aprendizado supervisionado pode ser melhor compreendido através da Figura 4.9. apresentada a seguir.

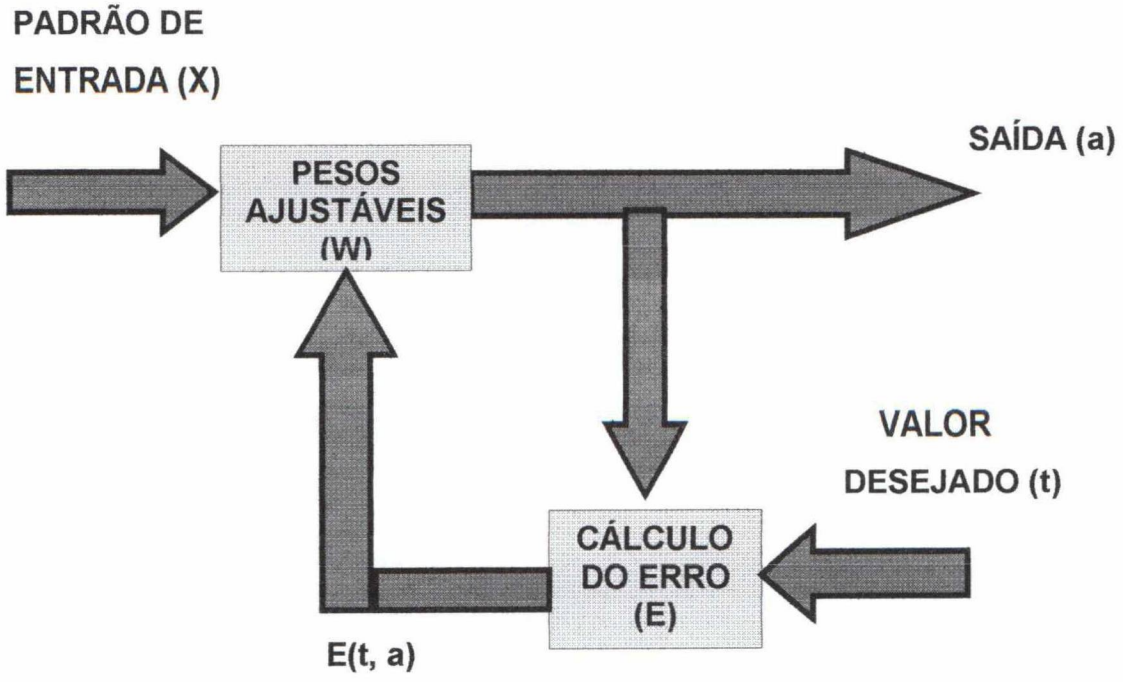


Figura 4.9. ESQUEMA DO APRENDIZADO SUPERVISIONADO

No aprendizado não-supervisionado, a rede "analisa" os conjuntos de dados apresentados a ela, determina algumas propriedades dos conjuntos de dados e "aprende" a refletir estas propriedades na sua saída. A rede utiliza padrões, regularidades e correlações para agrupar os conjuntos de dados em classes. As propriedades que a rede vai "aprender" sobre os dados pode variar em função do tipo de arquitetura utilizada e da forma de aprendizagem. Por exemplo, Mapa Auto-Organizável de *Kohonen,* Redes de *Hopfield* e Memória Associativa Bidirecional, são alguns métodos de aprendizago não-supervisionado (SIMEÃO, 1999).

O aprendizado não-supervisionado pode ser visualizado através do esquema apresentado pela Figura 4.10. a seguir.

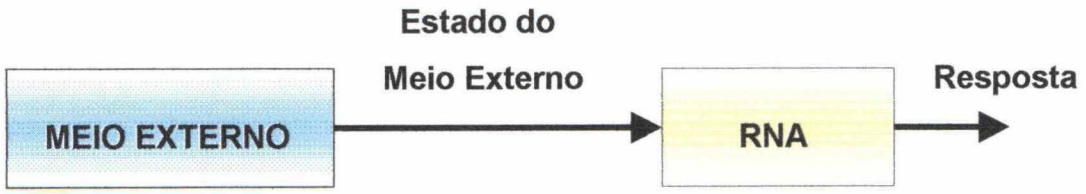


Figura 4.10. ESQUEMA DO APRENDIZADO NÃ0-SUPERVISIONADO

Pode-se classificar as redes neurais com relação a função de fluxo de dados. Por este aspecto as redes neurais classificam-se em: redes *feed-forward*, onde os dados propagam-se apenas unidirecionalmente, ou seja, apenas para a frente; ou redes *feedback* ou recursivas (recorrentes), se o fluxo de dados pode se dar entre todas as unidades.

4.5.4. Modelos de Redes Neurais

Muitos são os modelos de Redes Neurais. A seguir serão apresentados os modelos básicos: *Perceptron,* Redes Lineares e Redes de Múltiplas Camadas.

4.5.4.1. Perceptron

Conforme já mencionado, o perceptron foi proposto por Rosenblatt em 1959 e são redes do tipo *feed-forward* constituídas de unidades binárias, sendo o primeiro modelo de Redes Neurais. Este modelo de Rede Neural possui duas camadas, uma de entrada com n-elementos e outra de saída com número variável de elementos, dependendo do problema (KRÖSE e SMAGT, 1993)

Segundo GORNI (1993), este modelo foi proposto, numa tentativa de explicar a percepção visual da perspectiva de um neuro-fisiologista. O Perceptron pode ser visto como um instrumento de Reconhecimento de Padrões que não foi construído para reconhecer um conjunto específico de padrões, mas que tem alguma habilidade para aprender a reconhecer os padrões de um conjunto depois de um número finito de tentativas.

A Rede Perceptron de uma camada coloca limitações nos cálculos que um Perceptron pode executar. Os Perceptrons são treinados utilizando exemplos de comportamento correto. A regra de aprendizagem calcula as trocas nos pesos do Perceptron e *biases* dados em um vetor de entrada, $\underline{x}$, e o erro do Perceptron, E (STEINER, 1999).

Segundo STEINER (1999), o erro para um determinado padrão é simplesmente a diferença entre a resposta do neurônio, a, e o vetor alvo (ou saída desejada ao padrão), t (ou d_p). O vetor alvo, t, deve conter os valores "0" e "1". A

cada ajuste de pesos e *biases*, o Perceptron terá uma melhor chance de obter as saídas corretas, a = t (ou a = d_p), dados os vetores de entrada $\underline{x}$.

Assim, pode-se dizer, que o Perceptron é um modelo de rede neurais que aprende conceitos. Ele pode aprender a responder como verdadeiro (1) ou falso (0), "estudando" repetidamente os exemplos que lhe são apresentados através dos dados atribuídos aos neurônios da camada de entrada.

As Redes Perceptrons possuem muitas limitações, tais como:

- os valores de saída de um Perceptron podem ter somente 2 valores ("0" ou "1") devido a função de transferência que é utilizada (função passo);
- Perceptrons podem classificar somente conjunto de vetores linearmente separáveis;
- quando um vetor de entrada é muito maior ou muito menor que os demais, o processo de convergência é lento (GORNI, 1993).

Perceptrons são especialmente adequados para problemas simples de classificação de padrões.

Os Perceptrons, assim como qualquer rede neural, podem achar soluções diferentes ao iniciarem o processo de aprendizado de diferentes condições iniciais.

4.5.4.2. Redes Lineares

Estas redes diferem do Perceptron na função de transferência que, neste caso, é linear, permitindo que as saídas tomem qualquer valor entre "0" e "1" e não apenas os valores "0" e "1" como na função passo utilizada no Perceptron. Estas redes utilizam a regra de aprendizagem de *Widrow-Hoff*, também conhecida

como regra dos Mínimos Quadrados, para ajustar os pesos e *biases* de acordo com a magnitude dos erros, e não apenas pela sua presença (STEINER, 1999).

As redes lineares podem abordar dois tipos de situações-problemas:

- quando se deseja projetar uma rede linear para que ao se apresentar um conjunto de vetores de entrada, as saídas correspondam aos vetores alvo, desejado. Para cada vetor de entrada, calcula-se o vetor de saída da rede. A diferença entre o vetor de saída e o vetor alvo é o erro. O objetivo é achar valores para os pesos e *biases* da rede tal que a soma dos quadrados dos erros seja minimizada;
- quando se deseja projetar um sistema linear que possa responder a trocas do seu ambiente quando ele está operando. Tal sistema é chamado de sistema adaptativo. O primeiro trabalho neste campo foi feito por Widrow e Hoff que deram o nome de *ADALINE* para os elementos lineares adaptativos.

Como no Perceptron, as *biases* são úteis para provirem variáveis adicionais livres que podem ser ajustadas para auxiliarem na performance desejada para a rede.

Esta rede é algumas vezes chamada de *MADALINE* por conter muitas *ADALINES*.

É importante ressaltar o fato de que uma rede linear apresentar múltiplas camadas, não vai necessariamente resultar em uma rede mais poderosa; assim o fato de se usar uma única camada, não é uma limitação. Entretanto, Redes Lineares podem resolver problemas lineares, ou seja, contendo relações lineares entre as entradas e as saídas desejadas.

4.5.4.3. Redes de Múltiplas Camadas ou Redes em Avanço / Para Frente (*Feed-Forward*)

São modelos de redes que apresentam uma ou mais camadas entre as camadas de entrada e saída, chamadas camadas intermediárias. Este tipo de Rede Neural Artificial é o modelo mais utilizado atualmente; geralmente são treinadas através do algoritmo de Retropropagação (*Back-Propagation*).

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, seus pesos são 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, mas auxiliam no ajuste dos pesos da rede.

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 da entrada para a saída através das unidades intermediária.

Segundo CARVALHO (2000), a partir de extensões do Teorema de Kolmogoroff, são necessárias no máximo duas camadas intermediárias, com um número suficiente de unidades por camada, para se produzir quaisquer mapeamentos. Ainda, apenas uma camada intermediária é suficiente para aproximar qualquer função contínua.

4.5.5. Algoritmo de Retropropagação (*Back-Propagation)*

É o método mais utilizado para o treinamento de Redes Neurais e enquadra-se na aprendizagem supervisionada. Geralmente é aplicado a redes com múltiplas camadas do tipo em avanço (*feed-forward).*

Durante o treinamento com o algoritmo de Retropropagação (*Back-Propagation)*, a rede opera em uma seqüência de dois passos. Primeiro, um padrão 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. Se esta não estiver correta, o erro 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 modificados conforme o erro é retropropagado (CARVALHO, 2000).

As redes que utilizam o algoritmo de Retropropagação (*Back-Propagation)* trabalham com uma variação da regra delta, que é um método bastante simples de ajuste sináptico, que consiste no aprendizado de cada neurônio, isto é, quando um padrão é inicialmente apresentado à rede, ele produz uma saída, e após medir a diferença entre a resposta atual e a desejada, são realizados os ajustes apropriados nos pesos das conexões de modo a reduzir essa diferença. Este ajuste é realizado em função de um cálculo que 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.

Pode-se dizer que o algoritmo de Retropropagação (*Back-Propagation)* é a generalização da regra delta para funções de ativação não-lineares e redes multicamadas. Quando uma função linear é usada, a rede multicamadas não é

mais tão "poderosa" quanto à rede de uma única camada (KRÖSE e SMAGT, 1993).

Nestes casos, uma função de ativação amplamente utilizada é a função *sigmoidal.*

Segundo ADAMOWICZ (2000), cada vez que um padrão é apresentado à rede, os pesos são levemente modificados na direção requerida para produzir um menor erro na próxima vez que o mesmo padrão for apresentado. O grau de modificação de pesos é a taxa de aprendizagem. Quanto maior a taxa de aprendizagem, maiores as mudanças de pesos, e mais rapidamente será o processo de aprendizado. Oscilação ou não convergência podem ocorrer se a taxa de aprendizagem for muito grande. Assim, a taxa de aprendizagem é uma constante de proporcionalidade no intervalo [0, 1].

O treinamento das redes multicamadas com *(Back-Propagation)* pode demandar muitas iterações no conjunto de treinamento, resultando em um tempo de treinamento longo. Se for encontrado um mínimo local, o erro para o conjunto de treinamento estabiliza, estacionando em um valor maior que o aceitável. Uma maneira, que muitas vezes, faz aumentar o aprendizado sem levar o processo à oscilação é modificar a regra delta generalizada para incluir o termo *(momentum)*, uma constante que considera o efeito das mudanças passadas dos pesos na direção atual do movimento no espaço de pesos.

Desta forma, o termo *(momentum)* leva em consideração o efeito de mudanças anteriores de pesos na direção do movimento atual no espaço de pesos. O termo *(momentum)* torna-se útil em espaços de erro que contenham longas gargantas, com curvas acentuadas ou vales com descidas suaves (MENDES FILHO, 1997).

Em síntese, o algoritmo (*Back-Propagation)* é baseado no método gradiente descendente, que computa as derivadas parciais de uma função de erro, com relação ao vetor peso W de um certo vetor de entradas X. O treinamento da rede é dividido em duas fases principais: avante (*forward)* e retorno (*backward)*. A primeira etapa (*forward)* consiste na propagação dos estímulos apresentados da entrada para a saída. Esses estímulos fluem por toda a rede, recebendo a computação neural, camada por camada até gerarem a saída. A partir do resultado desejado *(target)*, calcula-se o erro na camada de saída. A segunda etapa (*backward)* ocorre em sentido contrário, onde o erro calculado, é retropropagado pelas camadas antecessoras, atualizando os pesos das conexões.

O algoritmo (*Back-Propagation)* pode ser modelado por funções matemáticas simples, conforme as descritas a seguir, quando sua topologia se enquadrar em um problema de classificação dicotômica (apenas uma unidade de saída é necessária) (STEINER, 1999).

Uma unidade *i* recebe os sinais de entrada e os agrega com base em uma função de entrada:

$$i_{p,i} = \sum_j w_{ij} x_{p,j} + \theta_i \qquad \text{p = 1, ... , (m + k)} \qquad [1]$$

$$\text{i = 1, ..., } k^*$$

$$\text{j = 1, ..., n}$$

onde:

$i_{p,i}$ = entrada da unidade i para o padrão p (sendo que o número total de padrões é igual a m + k);

i = número de unidades na camada escondida;

w_{ij} = conexão peso entre as unidades i e j;

$x_{p,j}$ = entradas (coordenadas) do padrão p;

θ_i = *bias* da unidade i.

Essa função de entrada gera um sinal de saída a_i para o padrão p, chamado de função de transferência:

$$a_{p,i} = \frac{1}{1+e^{-i_{p,i}}} \qquad [2]$$

Esses sinais de saída são então enviados para a única unidade da camada h, a qual os agrega em

$$i_{p,h} = \sum_i w_{hi} a_{p,i} + \theta_h, \qquad h = 1 \qquad [3]$$

gerando a saída:

$$a_{p,h} = \frac{1}{1+e^{-i_{p,h}}} \qquad [4]$$

Ainda, segundo STEINER (1999), na propagação (*forward)* os p = (m + k) padrões, descritos por suas coordenadas $x_{p,j}$, alimentam a rede conforme as equações de [1] a [4] já descritas. O valor de saída obtido para o padrão p, $a_{p,h}$, é comparado com o valor de saída desejado para o padrão p, d_p, calculando-se o erro quadrático:

$$E = \sum_{p=1}^{m+k} (d_{p,h} - a_{p,h})^2 / 2 \qquad [5]$$

O objetivo é minimizar E ajustado W de tal modo que todos os vetores de entrada sejam corretamente mapeados em suas correspondentes saídas. Então, o processo de aprendizagem pode ser visto como um problema de minimização com a função objetivo E definida no espaço de W.

A segunda fase, a propagação (*backward*), que envolve as equações de [6] a [9] a seguir, executa um gradiente descendente em W para localizar a solução que pode ser local ou eventualmente ótima global. A direção e magnitude Δw_{ij} pode ser calculada da seguinte forma:

- **Variação em w_{hi} 's**

$$\Delta_P w_{hi} = -\gamma \frac{\partial E_p}{\partial w_{hi}}$$

onde γ = taxa de aprendizagem, $0 < \gamma < 1$.

Desenvolvendo as derivadas, chega-se em:

$$\Delta_p w_{hi} = \gamma \; a_{p,h}(1 - a_{p,h}) \; a_{p,i}(d_p - a_{p,h}).$$

Considerando a situação atual t, para o padrão p, a troca de pesos ocorrida na situação (t - 1), para o padrão (p - 1), a fim de alcançar o mínimo mais rapidamente, tem-se:

$$\Delta_p w_{hi}(t) = \gamma \; a_{p,h} \; (1 - a_{p,h}) \; a_{p,i} \; (d_p - a_{p,h}) + \alpha \; \Delta_p w_{hi}(t-1) \qquad [6]$$

onde α = constante que determina o efeito na troca de pesos da iteração (t -1) (taxa *momentum)*

Então: $$w_{hi}(t) = w_{hi}(t-1) + \Delta_p w_{hi}(t) \qquad [7]$$

- **Variação em w_{ij} 's**

$$\Delta_P w_{hi} = -\gamma \frac{\partial E_p}{\partial w_{hi}}$$

Desenvolvendo as derivadas, chega-se em:

$$\Delta_p w_{ij} = \gamma \; (d_p - a_{p,h}) \; a_{p,h} \; (1 - a_{p,h}) \; w_{hi} \; a_{p,i} \; (1 - a_{p,i}) \; x_{p,j} \; ,$$

e ainda,

$$\Delta_p w_{ij}(t) = \gamma \, (d_p - a_{p,h}) \, a_{p,h} (1 - a_{p,h}) \, w_{hi} \, a_{p,i} \; (1 - a_{p,i}) \, x_{p,j} + \alpha \; \Delta_p w_{ij}(t-1) \qquad [8]$$

Então:

$$w_{ij}(t) = w_{ij}(t-1) + \Delta_p w_{ij}(t) \qquad [9]$$

4.5.6. Desenvolvimento de Aplicações

Segundo CARVALHO (2000), por ocasião da implementação das Redes Neurais à problemas reais, algumas etapas devem ser observadas, como as descritas nos itens a seguir.

4.5.6.1. Coleta e Separação de Dados em Conjuntos

O processo de desenvolvimento de Redes Neurais Artificiais inicia-se com a coleta de dados relativos ao problema. Esta tarefa requer uma análise cuidadosa sobre o problema para minimizar ambigüidades e erros nos dados. Além disso, os dados coletados devem ser significativos e cobrir amplamente o domínio do problema; não devem cobrir apenas as operações normais ou rotineiras, mas também as exceções e as condições nos limites do domínio do problema. Além disso, os dados são, geralmente, apresentados à rede em ordem aleatória para prevenção de tendências associadas à ordem de apresentação dos dados. Pode ser necessário pré-processar estes dados, através de normalizações, escalonamentos e conversões de formato para torná-los mais apropriados à sua utilização na rede.

Normalmente, os dados coletados são separados em dois conjuntos: conjunto de treinamento, que são os dados utilizados para o treinamento da rede e o conjunto de teste, que são dados utilizados para verificar a performance da rede sob condições reais de utilização. Além dessa divisão, pode-se usar também uma subdivisão do conjunto de treinamento, criando um conjunto de validação, utilizado para verificar a eficiência da rede quanto a sua capacidade de generalização durante o treinamento, podendo ser empregado como critério de parada do treinamento.

4.5.6.2. Configuração da Rede

O próximo passo é a definição da configuração da rede, que pode ser dividido em três etapas:

1) seleção do paradigma neural apropriado à aplicação, ou seja, que tipo de rede será utilizada;
2) determinação da topologia da rede a ser utilizada - o número de camadas escondidas e o número de unidades em cada camada;
3) determinação dos parâmetros do algoritmo de treinamento e funções de ativação. Este passo tem um grande impacto na performance do sistema resultante.

Normalmente estas escolhas são feitas de forma empírica. A definição da configuração de redes neurais é ainda considerada uma arte.

4.5.6.3. Treinamento da Rede Neural

A próxima etapa é o treinamento da rede. Nesta fase, seguindo o algoritmo de treinamento escolhido, serão ajustados os pesos das conexões. É importante considerar, nesta fase, alguns aspectos tais como a inicialização da rede, o modo de treinamento e o tempo de treinamento.

Uma boa escolha dos valores iniciais dos pesos da rede pode diminuir o tempo necessário para o treinamento. Com relação ao tempo de treinamento, vários fatores podem influenciar a sua duração, porém sempre será necessário utilizar algum critério de parada. Como critério de parada do algoritmo (*Back-Propagation),* em geral, é utilizado um número máximo de ciclos ou iterações.

4.5.6.4. Teste de uma Rede Neural

Depois de treinada, é preciso testar a Rede Neural. Durante esta fase, o conjunto de teste é utilizado para verificar a performance da rede com dados que não foram utilizados no treinamento. A performance da rede, calculada nesta fase, é uma boa indicação de sua performance real.

Devem ser considerados ainda outros testes como análise do comportamento da rede utilizando entradas especiais e análise dos pesos atuais da rede, pois se existirem valores muito pequenos, as conexões associadas podem ser consideradas insignificantes e assim serem eliminadas (*prunning*). De modo inverso, valores muito maiores que os outros poderiam indicar que houve (*over-training)* da rede (TATIBANA, 2000).

4.5.7. Parâmetros das Redes Neurais a serem considerados na Implementação Computacional

Ao se conceber uma Rede Neural Artificial do tipo Múltiplas Camadas com algoritmo de Retropropagação *(Back-Propagation)*, vários parâmetros devem ser considerados, como já visto, tais como: número de camadas de neurônios, número de neurônios em cada camada, taxa de aprendizagem e taxa de (*momentum).*

Uma Rede Neural deve conter no mínimo duas camadas de neurônios, ou seja, uma camada que se destina à entrada dos dados e outra que se destina à saída dos resultados. Entretanto, este tipo de rede apresentam uma utilidade muito limitada, pois atendem aos problemas em que os dados podem ser separados de forma linear.

O aumento do número de camadas de neurônios melhora o desempenho das redes neurais. Sua capacidade de aprendizado aumenta, o que se traduz na melhoria da precisão com que ela delimita as regiões de decisão (ADAMOWICZ, 2000).

Embora a necessidade de que haja pelo menos uma camada oculta na Rede Neural seja praticamente um ponto pacífico, há considerável controvérsia quanto ao número necessário de camadas ocultas.

Após definir o número de camadas a serem utilizadas na Rede Neural, a próxima etapa é definir o número de neurônios por camada.

Na camada de entrada deve haver um número de neurônios igual ao número de variáveis a serem fornecidos à rede. Eventualmente, uma variável de entrada pode ser sub-dividida em vários neurônios, segundo um esquena binário, o que pode melhorar o desempenho.

A camada de saída deve conter um número de neurônios igual ao número de variáveis que se deseja calcular. No caso de modelos classificatórios, pode-se utilizar um neurônio para cada item de classificação ou utilizar uma representação mais compacta, empregando-se técnicas binárias para diminuir o número de neurônios.

Contudo, o uso de representação binária na camada de saída aumenta a carga de trabalho da camada oculta, obrigando a um aumento do número de neurônios dessa camada ou mesmo a adição de uma camada oculta suplementar para que a Rede Neural mantenha o mesmo nível de desempenho.

A literatura sugere vários critérios para a escolha do número ótimo de neurônios das camadas ocultas, porém, de forma geral, para redes pequenas, onde o número de neurônios na camada de saída é maior do que o da entrada, sendo que a média geométrica entre o número de neurônios nas camadas de entrada e saída (ou seja, $\sqrt{(is)}$, onde i = número de neurônios da camada de entrada e s = número de neurônios na camada de saída) é uma boa estimativa para o número de neurônios da camada oculta. Por outro lado, quanto mais complexo for o relacionamento entre as variáveis de entrada e de saída, maior deverá ser o número de neurônios na camada oculta (GORNI,1993).

Outra questão a ser tratada por ocasião da implementação de uma Rede Neural diz respeito aos valores dos parâmetros γ (taxa de aprendizagem) e α (*momentum*) contidos no algoritmo (*Back-Propagation).*

Alguns autores sugerem que a taxa de aprendizagem da Rede Neural, definida por um coeficiente de aprendizado γ ($0 < \gamma < 1$), deve ser alta no início do treinamento e decline gradativamente à medida que ele evolui (GORNI, 1993). Isto deve proporcionar rapidez na convergência do treinamento, estabilidade e resistência ao aparecimento de mínimos locais.

Para propósitos práticos escolhe-se uma taxa de aprendizagem que seja tão grande quanto possível sem levar à oscilação. Uma maneira de evitar oscilação é fazer a atualização dos pesos depender da atualização de pesos passadas adicionando o termo (*momentum)*, conforme já mencionado. Quando o termo (*momentum)* é acrescentado e a taxa de aprendizagem é pequena, a convergência da rede é lenta; quando o termo (*momentum)* não é considerado e a taxa de aprendizagem é alta, o mínimo nunca é alcançado, porque ocorrem oscilações e, finalmente, quando a taxa de aprendizagem é alta, mas o termo (*momentum)* é considerado, o mínimo é alcançado rapidamente (GORNI, 1993).

Um coeficiente de aprendizado inicial muito alto, dados com magnitude incompatível com a função de ativação dos neurônios ou uma rede mal dimensionda, podem fazer com que não haja convergência já na fase de treinamento (HADJIPROCOPIS, 1993).

4.6. ÁRVORES DE DECISÃO

Neste item, será feita uma abordagem a outra técnica de *Data Mining* utilizada na resolução deste trabalho.

4.6.1. Introdução

As Árvores de Decisão são uma evolução das técnicas de aprendizado de máquina (*machine learning).*

Foi um professor da Universidade de Sydney, Austrália, Ross Quinlan, que desenvolveu a tecnologia que permitiu o aparecimento das Árvores de Decisão. Muitas pessoas na indústria de *Data Mining* consideram Quinlan como o "pai das Árvores de Decisão". A contribuição de Quinlan foi a elaboração de um novo algoritmo chamado ID3, desenvolvido em 1983. O ID3 e suas evoluções (ID4, ID6, C 4.5, See 5) são muito bem adaptadas para usar em conjunto com as árvores de decisão, na medida em que eles produzem regras ordenadas pela importância. Essas regras são, então, usadas para produzir um modelo de árvore de decisão dos fatos que afetam os ítens de saída (DWBRASIL, 2002).

Árvores de Decisão fazem parte do contexto de *Data Mining*, dos métodos de classificação. As Árvores de Decisão são, quase sempre, usadas em conjunto

com a tecnologia de Indução de Regras, mas são únicas no sentido de apresentar os resultados da Indução de Regras num formato com priorização. Então, o atributo mais importante é apresentado na árvore como o primeiro nó, e os atributos menos relevantes são mostradas nos nós subseqüentes. As vantagens principais das Árvores de Decisão são que elas "tomam decisões" levando em consideração os atributos que são mais relevantes, além de serem 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.

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.* Um problema complexo é decomposto em sub-problemas mais simples. Recursivamente a mesma estratégia é aplicada a cada sub-problema (GAMA, 2000).

A capacidade de discriminação de uma Árvore advém das seguintes características:

- divisão do espaço definido pelos atributos em sub-espaços;
- a cada sub-espaço é associada uma classe.

Segundo GARCIA (2000), as Árvores de Decisão consistem de:

- nodos (nós) que representam os atributos;

- de arcos (ramos), provenientes destes nodos e que recebem os valores possíveis para estes atributos (cada ramo descendente corresponde a um possível valor deste atributo) e

- de 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.

Na Árvore de Decisão cada nó deve estar associado a um atributo que é o mais informativo entre os atributos ainda não considerados no caminho desde a raiz.

A Figura 4.11 apresenta um exemplo de árvore de decisão, explicada mais detalhadamente mais adiante.

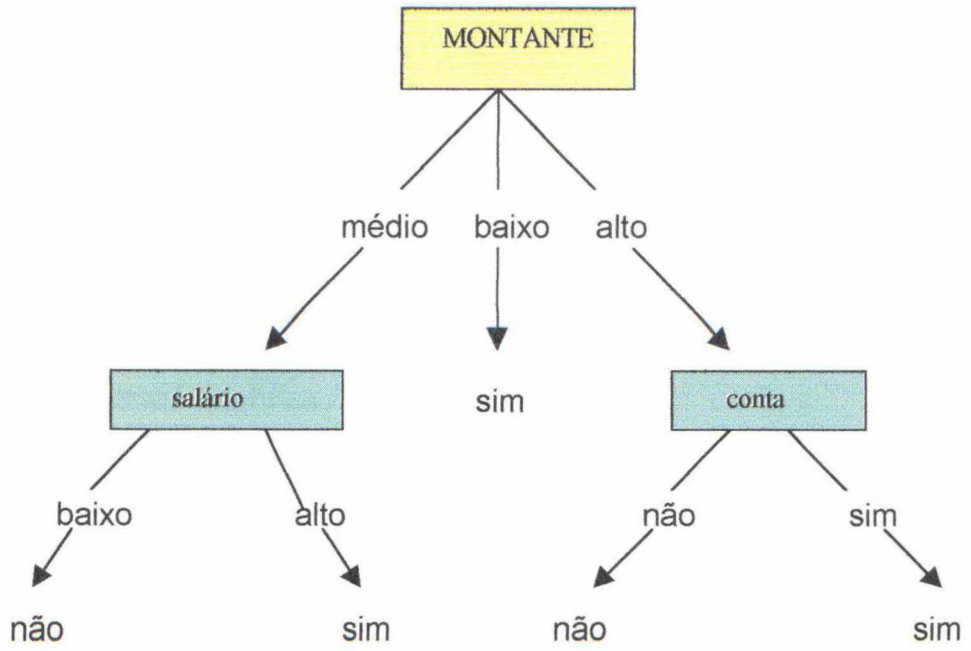


Figura 4.11. Exemplo de uma Árvore de Decisão

Neste exemplo são trabalhados dados que relatam as condições propícias de uma pessoa receber ou não um empréstimo. Tem-se então 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 ser baixo ou alto e o atributo conta pode ser "sim" ou "não". Alguns dados são exemplos positivos de uma classe "sim", ou seja, os requisitos exigidos a uma pessoa, por um banco, são satisfatórios à concessão de um empréstimo, e outros são negativos, onde os requisitos exigidos não são satisfatórios à concessão de um empréstimo. Classificação, neste caso, é a construção de uma estrutura de árvore, que pode ser usada para classificar corretamente todos os objetos do conjunto (BRADZIL, 1999).

Após a construção de uma Árvore de Decisão é possível derivar regras. Essa transformação da Árvore de Decisão em regras, geralmente é feita no intuito de facilitar a leitura e a compreensão humana. 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 das Árvores de Decisão tenderem a crescer muito, de acordo com algumas aplicações, elas são muitas vezes substituídas pelas regras. Isto acontece em virtude das regras poderem ser facilmente modularizadas. Uma regra pode ser compreendida sem que haja a necessidade de se referenciar outras regras (INGARGIOLA, 1996).

Com base na árvore de decisão apresentada na Figura 4.12, pode-se exemplificar a derivação de regras. Dois exemplos de regras obtidas a partir desta árvore são mostrados a seguir.

- **SE** montante = médio **e** salário = baixo
  ENTÃO classe = não.

- **SE** montante = médio **e** salário = alto
  ENTÃO classe = sim.

4.6.2. Objetivos das Árvores de Decisão

Uma Árvore de Decisão tem a função de particionar recursivamente um conjunto de treinamento, até que cada subconjunto obtido deste particionamento contenha casos de uma única classe, obtendo-se assim um modelo que servirá para futuras classificações (QUINLAN, 1993).

Para atingir esta meta, a técnica de Árvores de Decisão examina e compara a distribuição de classes durante a construção da árvore. O resultado obtido, após a construção de uma Árvore de Decisão, são dados organizados de maneira compacta, que são utilizados para classificar novos casos (BRAZDIL, 1999).

Sintetizando, o objetivo é gerar os valores categóricos de um atributo chamado "classe".

4.6.3. Construção de uma Árvore de Decisão

O processo de construção de uma Árvore de Decisão inicia-se a partir de um conjunto de treinamento, que contém exemplos com classes previamente conhecidas (dados históricos). O conjunto de treino deve conter exemplos positivos e negativos de uma classe.

Seja T o conjunto de treinamento, composto pelas classes {c1, c2, ..., cn}. A idéia é dividir T em subconjuntos que contenham casos, todos pertencendo a uma mesma classe cj. Essa divisão é feita baseada em um atributo que possua valores mutuamente exclusivos {v1, v2, v3, ..., vn}. O conjunto T é particionado em subconjuntos T1, T2, T3, ..., Tn, onde Ti contém todos os casos com valores vi. A Árvore de Decisão para T consiste de um nó de decisão identificando o teste para o atributo e um galho para cada valor do atributo. Recursivamente, cada subconjunto Ti é visto como T até que todos os elementos de Ti pertençam a uma mesma classe cj (QUINLAN, 1993).

Para gerar uma árvore de decisão com uma alta taxa de predição é necessário fazer a escolha correta dos atributos que serão usados como teste no agrupamento dos casos. Estes testes devem gerar uma árvore com o menor número possível de subconjuntos, fazendo com que cada folha da árvore

contenha um número significativo de casos. O ideal é escolher os testes de modo que a árvore final seja a menor possível.

Como analisar todas as possibilidades possíveis seria algo absurdo, foram desenvolvidos vários métodos aplicados na escolha dos atributos e dos testes a serem utilizados, porém todos concordam em dois pontos: uma divisão que mantém as proporções de classes em todas as partições é inútil e uma divisão onde em cada partição todos os exemplos são da mesma classes tem utilidade máxima. Uma vez feita a escolha, as outras possibilidades não são mais exploradas.

Para melhor esclarecer os critérios que levam à escolha de um atributo, faz-se necessário a introdução de dois conceitos: **entropia** e **ganho de informação.**

4.6.3.1. Entropia

Entropia é a medida que indica a homogeneidade dos exemplos contidos em um conjunto de dados. Ela permite caracterizar a "pureza" (e impureza) de uma coleção arbitrária de exemplos (OSÓRIO, 2000).

Dado o conjunto S, contendo exemplos '+' e '-' que definem o conceito a ser aprendido, a entropia relativa dos dados deste conjunto S é indicada por:

$$\text{Entropia}(S) = - P_{(+)} \cdot \log_2 P_{(+)} - P_{(-)} \cdot \log_2 P_{(-)} \qquad [10]$$

onde:

$P_{(+)}$ = Proporção entre os exemplos positivos e o total de exemplos do conjunto. (número de casos positivos / número total de casos).

$P_{(-)}$ = Proporção entre os exemplos negativos e o total de exemplos do conjunto. (número de casos negativos / número total de casos).

É assumido que: $0 \cdot \log_2 0 = 0$, por definição.

A equação [10] apresentada é usada para calcular a entropia levando-se em conta duas classes. Generalizando para "N" Classes, a equação fica:

$$\text{Entropia (S)} = -\sum_{i=1}^{N} P_i \log_2 P_i \qquad [11]$$

A entropia (S) tem máximo valor para ($\log_2 P_i$) se $P_i = P_j$ para qualquer $i \neq j$ (existem tantos elementos positivos como negativos) e a entropia (S) = 0, se existe um i tal que $p_i = 1$ (todos os elementos são da mesma classe). Por exemplo: se P for dada por (0,5; 0,5), então entropia (P) é igual a 1; se P for dada por (0,67; 0,33) então entropia (P) será 0,92; se P for dada por (1,0) então entropia (P) será 0.

4.6.3.2. Ganho de Informação (Critério *GAIN*)

Segundo OSÓRIO (2000), o ganho de informação é a medida que indica o quanto um dado atributo irá separar os exemplos de aprendizado de acordo com a sua função objetivo (classes).

O ganho de informação é a redução esperada no valor da Entropia devido à ordenação do conjunto de treino segundo os valores do atributo escolhidos (ANTUNES, 2000).

GAIN (S, A) = Redução esperada na entropia de S, causada pelo particionamento dos exemplos em relação a um atributo escolhido (A).

$$\text{Gain (S, A)} = \text{Entropia (S)} - \sum_{v=1}^{N} \frac{|S_v|}{|S|} \cdot \text{Entropia}(S_v) \qquad [12]$$

onde:

A = Atributo considerado;
N = Número de valores possíveis que este atributo pode assumir;
S_v = Sub-conjunto de S onde o atributo A possui o valor V.

O método de particionamento recursivo para geração das Árvores de Decisão, que subdivide o conjunto de casos de treinamento até que cada subconjunto em cada partição contenha casos de uma única classe ou até que nenhum outro teste ofereça qualquer melhora, pode gerar árvores complexas que acabam perdendo o seu poder de predição. Faz-se necessário então, adotar algumas medidas para tornar árvores complexas em árvores mais simples (QUINLAN, 1993).

Existem dois caminhos pelos quais este particionamento recursivo pode ser modificado para produzir árvores mais simples: decidindo não continuar a dividir o conjunto de dados de treinamento ou removendo retrospectivamente alguma estrutura já construída pelo método.

O primeiro caminho pode causar o término da divisão antes que o benefício das divisões subseqüentes se tornem evidentes.

Na segunda alternativa, o processo de *dividir-para-conquistar* segue até o fim e então, a árvore é "podada". Este processo é mais lento, mas muito mais seguro. O processo de poda irá, invariavelmente, causar a mistura nas classes de alguns casos de treinamento.

Pode-se dizer que uma das maiores motivações para podar Árvores de Decisão, é no sentido de se evitar o ajuste demais / sobreajuste (*overfitting)* da árvore aos dados. Neste caso, a árvore poderia se ajustar a pecularidades dos dados, que talvez não ocorram em dados ainda não vistos (FREITAS, 2000).

Ainda, segundo FREITAS (2000), deve-se ter cuidado para que a poda não seja muito agressiva, a fim de não gerar um sub-ajustamento (*underfitting)* da árvores aos dados.

A realização da "poda" ou simplificação das Árvores de Decisão é baseada em "erros".

O algoritmo de simplificação começa pelo nível mais baixo da árvore e vai examinando cada nó. Para cada nó ele verifica se a deleção da subárvore em favor de uma folha ou a substituição pelo galho mais freqüente deste nó irá diminuir a predição de erro; caso positivo a operação é realizada.

Pode-se descrever um algoritmo básico de (*pruning)* poda, através dos seguintes passos:

- Percorrer a árvore em profundidade;
- Para cada nó de decisão calcular:
  . o erro no nó;
  . a soma dos erros nos nós descendentes;

- Se o erro no nó é menor ou igual à soma dos erros dos nós descendentes o nó é transformado em folha;
- Eliminar os nós descendentes.

Ainda resta um problema: como esta predição de erro pode ser encontrada? Se o conjunto de casos de treinamento for usado para calcular esta predição, ao se tentar simplificar a árvore, a predição de erros irá aumentar, pois a árvore absorveu o padrão existente nesse conjunto, e com isso não irá ocorrer simplificação. Para solucionar este problema pode-se separar um pequeno conjunto de casos e não usá-los na geração da árvore. Assim a predição de erro não seria influenciada, pois ela seria feita com um conjunto desconhecido de casos. Apesar de ser uma ótima solução, esta técnica traria outros problemas quando o conjunto de dados fosse pequeno (QUINLAN, 1993).

Existe um artifício para contornar o problema da árvore ter absorvido o conhecimento do conjunto de casos de treinamento. A técnica é chamada de simplificação pessimista e consiste em aumentar a taxa de erro ocorrida em cada folha e então inserir os casos novamente para encontrar a predição de erros. O algoritmo C4.5 faz uso desse artifício (AURORA, 2000).

4.6.4. Generalidade sobre Árvores de Decisão

Muitos são os algoritmos de classificação que elaboram Árvores de Decisão. Não há uma forma de determinar qual é o melhor algoritmo, sendo que um algoritmo pode ter melhor desempenho em determinada situação e outro pode ser mais eficiente em outros tipos de situações.

O algoritmo ID3 foi um dos primeiros algoritmos de Árvore de Decisão, tendo sua elaboração baseada em sistemas de inferência e em conceitos de sistemas

de aprendizagem. Logo após foram elaborados diversos algoritmos, sendo os mais conhecidos: C4.5, CART (*Classification and Regression Trees*), CHAID (*Chi Square Automatic Interaction Detection*), entre outros (GARCIA, 2000).

A utilização de Árvores de Decisão apresenta as seguintes vantagens:

- Não assume nenhuma distribuição particular para os dados;
- As características ou atributos podem ser categóricos (qualitativos) ou numéricos (quantitativos);
- Pode construir modelos para qualquer função desde que o número de exemplos de treino seja suficiente;
- Elevado grau de interpretabilidade.

Após a construção de uma Árvore de Decisão é importante avaliá-la. Esta avaliação é realizada através da 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, se estimar a proporção de erros e acertos ocorridos na construção da árvore (BRAZDIL, 1999).

CAPÍTULO V

5. IMPLEMENTAÇÀO DAS TÉCNICAS AO PROBLEMA E ANÁLISE DOS RESULTADOS

5.1. INTRODUÇÃO

Neste capítulo será descrita a forma como foram implementadas ao problema, duas das técnicas de *Data Mining*: Árvores de Decisão e Redes Neurais, bem como a comparação entre elas no que diz respeito à sua eficiência.

5.2. ÁRVORES DE DECISÃO

Na implementação da técnica Árvores de Decisão optou-se por utilizar o *software* computacional WEKA *(Waikato Environment for Knowledge Analysis)*, tendo em vista sua praticidade de utilização, bem como ser um *software* de domínio público estando disponível em: http://www.cs.waikato.ac.nz/ml/weka.

O *software* WEKA é formado por um conjunto de algoritmos de diversas técnicas para resolver problemas concretos de *Data Mining.* O WEKA está implementado em linguagem Java e foi desenvolvido no meio acadêmico da Universidade de Waikato, na Nova Zelândia, em 1999.

Os algoritmos que compõem o WEKA são alocados em forma de pacotes como, por exemplo, os listados a seguir.

- *package weka.associations*
- *package weka.attributeSelection*
- *package weka.classifiers*
- *package weka.classifiers.j48*
- *package weka.classifiers.neural*
- *package weka.clusterers*
- *package weka.core*
- *package weka.estimators*
- *package weka.filters*
- *package weka.gui*
- *package weka.gui.experiment*
- *package weka.gui.explorer*
- *package weka.gui.streams*
- *package weka.gui.treevisualizer*
- *package weka.gui.visualize*

Esses pacotes são ainda subdivididos em vários outros subpacotes que procuram se ater a atividades mais específicas de *Data Mining*, como: classificação, associação, estimativas, filtragens, interfaces gráficas e outros.

5.2.1. Formato do Arquivo de Entrada do WEKA

O *software* WEKA utiliza o padrão ARFF para seus arquivos de entrada, independentemente do algoritmo utilizado.

O padrão ARFF é utilizado para representar uma série de dados que consistem em exemplos independentes. As especificações dos atributos em linhas de ARFF permitem que a série de dados sejam checadas quanto à sua consistência de forma automática pelos programas que lêem as linhas de ARFF (WITTEN e FRANK, 2000).

Assim, para que se possa aplicar os dados a qualquer algoritmo do pacote WEKA, se faz necessário primeiramente que estes sejam convertidos para o formato ARFF.

Num arquivo com extensão ARFF contendo dados num formato apto a ser processado pelo WEKA, tem-se:

- opcionalmente, linhas começando com o símbolo %; significa que a linha é um comentário, não tendo validade junto ao processamento realizado pelos algoritmos;

- a primeira linha válida indica o nome da relação a encontrar (@relation nome_da_relação);

- as linhas seguintes devem listar todos os atributos, onde deve-se definir o tipo do atributo e os valores que ele pode representar. Neste último caso, os valores devem estar entre "{ }" e separados por vírgulas;

- no próximo bloco de informações, após uma linha de indicação (@data), vêm as instâncias, ou seja, os registros a serem minerados com o valor dos atributos para cada instância separado por vírgula; a ausência de um item em um registro deve ser atribuída pelo símbolo "?".

A Figura 5.1. abaixo, apresenta o arquivo "pessoajurídica" , extensão ARFF, contendo exemplos dos dados das 339 empresas analisadas neste trabalho.

```
@relation pessoajuridica

@attribute restrições {sim, nao}
@attribute restrições5anos {sim, nao}
@attribute tempoconta real
@attribute atividade {comercio, industria, servico}
@attribute tempoatividade {1, 2, 3, 4, 5}
@attribute funcionarios real
@attribute sede {proprio, alugado, cedido}
@attribute bairro {centro, outro}
@attribute clientes {fisica, juridica, misto}
@attribute faturamento real
@attribute outrobanco {sim, não}
@attribute imóveis real
@attribute móveis real
@attribute seguroempresa {sim, não}
@attribute aplicações {1, 2, 3, 4, 5}
@attribute vendasprazo {1, 2}
@attribute experiênciacredito {1, 2, 3}
@attribute historicoc/c {1, 2, 3, 4}
@attribute sóciosrestrições {sim, não}
@attribute sóciosrestri5anos {sim, não}
@attribute sociedadecônjuges {sim, não}
@attribute imóveissócios real
@attribute móveissócios real
@attribute risco {A, B, C, D, E}
@attribute resultado {S, N}

@data
nao,nao,17,comercio,5,5,alugado,centro,fisica,12277,não,0,0,não,5,2,1,2,n
ão,sim,sim,0,0,E,S
nao,nao,175,industria,5,8,proprio,outro,juridica,157855,não,115750,0,não,
5,2,2,1,não,não,sim,0,0,A,S
....
```

Figura 5.1. Arquivo pessoajurídica.arff

Os dados constantes dessa Figura 5.1., obedecem os critérios descritos no Capítulo II, item 2.2, ou seja, cada linha de atributo corresponde as informações (variáveis) que estão sendo analisadas neste trabalho.

Assim, por exemplo, no atributo resultado a informação "S" corresponde a "Adimplente" e a informação "N" corresponde a "Inadimplente".

5.2.2. Implementação computacional

Na implementação da técnica de Árvores de Decisão, foi utilizada a base de dados deste trabalho, composta de informações de 339 empresas (266 adimplentes e 73 inadimplentes), utilizando na execução o algoritmo de classificação J48 (C4.5 *release* 8).

Foram realizados 8 conjuntos de testes no total. O primeiro teste realizado contemplou informações de todas as 339 empresas. Nos demais, os dados foram separados em dois conjuntos: um contendo informações de 306 empresas (241 adimplentes e 65 inadimplentes) utilizado para geração da Árvore de Decisão e outro contendo informações de 33 empresas (25 adimplentes e 8 inadimplentes) que foi utilizado para testar a Árvore gerada.

Em cada um dos testes realizados, com exceção do primeiro, os conjuntos foram gerados de forma aleatória, a fim de se evitar qualquer tipo de indução de resultados.

Os resultados obtidos constam da Tabela 5.2. a seguir.

ÁRVORES DE DECISÃO						
TESTES	CONJUNTO TREINAMENTO			CONJUNTO TESTES		
	ADIM-PLENTES	INADIM-PLENTES	ERRO PERCENTUAL	ADIM-PLENTES	INADIM-PLENTES	ERRO PERCENTUAL
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%

Tabela 5.2. RESULTADOS OBTIDOS COM A TÉCNICA ÁRVORES DE DECISÃO

Com base nos dados apresentados na Tabela 5.2., observa-se que o Teste 5 apresentou o melhor resultado 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, constante do ANEXO III deste trabalho, apresentou a seguinte performance: 282 registros (instâncias) classificados

corretamente e 24 registros classificados de forma incorreta correspondendo, respectivamente, a 92,16% e 7,84% do total de registros.

Dos 282 registros classificados corretamente, 235 referem-se a empresas adimplentes e 47 referem-se a empresas inadimplentes. Dos 24 registros classificados incorretamente, 6 são de empresas adimplentes e 18 são de empresas inadimplentes.

A Árvore de Decisão gerada pelo Teste 5, apresentou um número de folhas igual a 40, ou seja, pode-se extrair 40 regras do tipo SE-ENTÃO, conforme descrito no capítulo IV, item 4.3.1. , algumas das quais relacionadas a seguir:

- SE tempo de conta > 25 meses
  ENTÃO adimplente.

- SE tempo de conta ≤ 25 meses, sócios não tem restrições, a empresa possui risco A
  ENTÃO adimplente.

- SE tempo de conta ≤ 25 meses, sócios não tem restrições, a empresa possui risco B, sócios não tiveram restrições baixadas nos últimos 5 anos e a sociedade é entre cônjuges
  ENTÃO adimplente.

- SE tempo de conta ≤ 25 meses, sócios não tem restrições, a empresa possui risco C e a empresa teve restrições baixadas nos últimos 5 anos
  ENTÃO inadimplente.

- SE tempo de conta ≤ 25 meses, sócios não tem restrições, a empresa possui risco C e a empresa não teve restrições baixadas nos últimos 5

anos, principais clientes são pessoas físicas e não possui seguro empresarial
ENTÃO inadimplente.

5.3. 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 a utilizada na técnica de Árvores de Decisão, ou seja, foram realizados 8 conjuntos de testes no total. O primeiro teste realizado contemplou informações de todas as 339 empresas. Nos demais, os dados foram separados em dois conjuntos: um contendo informações de 306 empresas (241 adimplentes e 65 inadimplentes) utilizado para treinamento da Rede Neural e outro contendo informações de 33 empresas (25 adimplentes e 8 inadimplentes) que foi utilizado para testar a Rede. Na realização de cada teste, os registros dos conjuntos de treinamento e de testes foram os mesmos utilizados para os testes aplicados à técnica Árvores de Decisão.

Os treinamentos foram feitos através de uma rede de múltiplas camadas, usando o algoritmo *Back-Propagation* padrão, variando os seguintes parâmetros:

- quantidade de iterações (ou ciclos): em cada conjunto de teste, o conjunto utilizado para treinamento da Rede foi submetido as seguintes quantidades de iterações: 100, 1.000, 2.000, 4.000, 6.000, 8.000 e 10.000, portanto o número máximo de iterações foi limitado a 10.000 em cada teste realizado;

- quantidade de neurônios intermediários (ou escondidos) da rede: em cada teste realizado, a Rede foi treinada primeiramente sem a camada intermediária e nos demais testes utilizando respectivamente, 2, 4, 6, 8 e 10 neurônios na camada intermediária.

- Para cada teste realizado foi utilizado um conjunto de pesos iniciais aleatório, num total de 48 conjuntos (8 testes X 6 quantidades de neurônios na camada intermediária).

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 trabalho igual a 24 e a quantidade de neurônios na camada de saída é igual a 1.

5.3.1. Formato de Entrada de Dados no MATLAB

Para a utilização do *software* MATLAB na implementação computacional das Redes Neurais, foi preciso alocar a base de dados deste trabalho em uma matriz P de entrada de dados na Rede, contendo no primeiro teste 339 linhas e nos demais 306 (que correspondem ao número de registros de empresas utilizados em cada teste) e 24 colunas (que correspondem ao número de informações de cada empresa).

Nos primeiros testes realizados, a matriz P foi "alimentada" com os dados exatamente como constam do ANEXO II deste trabalho, porém os resultados obtidos no treinamento não foram satisfatórios.

Deste modo, numa tentativa de melhorar a performance da Rede, optou-se por tabular os dados referentes a faturamento e valores dos bens imóveis e móveis, através de faixas de valores. Cada faixa compreendeu um intervalo de R$ 30.000,00, assim, por exemplo, um valor de R$ 324.000,00 foi substituído pela informação 11.

Após essa adaptação aos dados, a nova matriz P foi utilizada para o treinamento da Rede.

As saídas (adimplente = 1 e inadimplente = 0) foram alocadas em uma matriz T, cujo número de linhas foi igual ao da matriz P e cujo número de colunas foi igual a 1.

O treinamento da Rede Neural através do MATLAB é feito utilizando as funções descritas a seguir.

- **FUNÇÃO INITFF**: o treinamento é inicializado através desta função, que examina a matriz com vetores de entrada P e as funções de transferência de cada camada e retorna os pesos W e polarizações (*biases)* b para cada camada.
  Exemplo: Rede Neural com 8 neurônios na camada escondida, 1 neurônio na camada de saída e função de transferência sigmoidal (logsig).
  [w1,b1,w2,b2]=initff(P,8,'logsig',1,'logsig'), onde w1 e b1 são pesos e bias da 1ª camada; w2 e b2 são pesos e bias da 2ª camada e logsig é a função de ativação sigmoidal.

- **PARÂMETROS TP**: através deste comando são especificados os parâmetros da Rede Neural: o número de iterações compreendido entre cada apresentação das respostas, o número de iterações, o valor do erro esperado e variação da taxa de aprendizagem.

Exemplo: tp=[1 100 0.01 0.01], neste exemplo o número 1 significa que a resposta deve ser mostrada a cada iteração; o número 100 é o número de iterações; o número 0.01 é o valor do erro a ser atingido; o número 0.01 é a variação da taxa de aprendizagem.

- **FUNÇÃO TRAINBP**: usada para treinar a rede através do algoritmo *Back-Propagation* (BP).
  Exemplo: [w1,b1,w2,b2,te,tr]=trainbp(w1,b1,'logsig',w2,b2,'logsig',P,T,tp), onde: w1, b1, w2, b2,logsig e tp já foram definidos anteriormente. O te indica o número de iterações e o tr indica o erro de cada iteração.

- **FUNÇÃO SIMUFF**: esta função examina a entrada P da rede, os pesos W, as polarizações b, a função de transferência para até três camadas e apresenta o retorno através da saída "a2".
  Exemplo: a2=simuff(P,w1,b1,'logsig',w2,b2,'logsig'). Este comando fornece os resultados para cada um dos padrões.

Para efetuar a verificação dos resultados dos conjuntos de Testes das Redes Neurais, foi elaborado um programa através do aplicativo Excel.

Os melhores resultados encontrados em cada conjunto de testes constam da Tabela 5.3. a seguir.

REDES NEURAIS							
TESTES	QUANTIDADE NEURÔNIOS CAMADA ESCONDIDA	CONJUNTO TREINAMENTO			CONJUNTO TESTES		
		ADIM-PLENTES	INADIM-PLENTES	ERRO PERCENTUAL	ADIM-PLENTES	INADIM-PLENTES	ERRO PERCENTUAL
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%

Tabela 5.3. RESULTADOS OBTIDOS COM A TÉCNICA REDES NEURAIS

Os resultados apresentados na Tabela 5.3. acima, foram obtidos após 10.000 iterações em cada conjunto de treinamento.

Analisando a Tabela 5.3., observa-se que o melhor resultado obtido com a técnica de Redes Neurais foi o encontrado através do Teste 6. O mesmo apresenta na fase de treinamento, onde o objetivo é fazer com que a Rede aprenda através dos dados a ela apresentados, a distinguir empresas adimplentes

das inadimplentes, um erro de 2,28% e na fase de teste, onde a Rede treinada é utilizada como base de conhecimento, ou seja, como instrumento de apoio à decisão, auxiliando na classificação de uma empresa qualquer em adimplente ou inadimplente, o erro apresentado foi de 3,03%.

5.4. ANÁLISE DOS RESULTADOS OBTIDOS

Neste presente capítulo foram apresentadas as implementações computacionais dos dados do problema e os testes feitos empregando as técnicas de Árvores de Decisão e Redes Neurais para a discriminação de clientes pessoas jurídicas como adimplentes ou inadimplentes.

Os resultados obtidos, do ponto de vista do percentual de erros apresentados em ambas as técnicas se encontram sintetizados na Tabela 5.4. a seguir.

	MÉDIAS DOS ERROS	
	ÁRVORES DE DECISÃO	REDES NEURAIS
FASE TREINAMENTO	11,49%	4,09%
FASE DE TESTE	28,13%	9,96%

Tabela 5.4. MÉDIA DOS ERROS VERIFICADOS NAS TÉCNICAS DE ÁRVORES DE DECISÃO E REDES NEURAIS

A performance da técnica de Redes Neurais foi melhor do que a apresentada pela técnica de Árvores de Decisão, conforme pode ser verificado na Tabela 5.4., porém do ponto de vista do usuário das ferramentas utilizadas, uma pequena vantagem na utilização da técnica Árvores de Decisão, no sentido de que a mesma apresenta seus resultados com uma aparência de fácil compreensão, através de regras, detalhando quais das informações a respeito das empresas analisadas foram mais relevantes na classificação.

Vale salientar que, na apresentação dos dados qualitativos para a Rede Neural, os mesmos precisaram ser transformados em valores numéricos, como descrito no Capítulo II, item 2.2. Já na técnica Árvores de Decisão, foi possível a entrada de dados com valores quantitativos e qualitativos, simultaneamente.

Com o uso das técnicas aqui apresentadas será possível proceder a análise de uma nova proposta de concessão de crédito, com uma margem maior de segurança, pois as mesmas sinalizam para que:

- mais merecedores de crédito recebam crédito, aumentando os lucros;
- mais não merecedores de crédito tenham seus pedidos de crédito negados ou reduzidos, ou ainda, exigidas garantias adicionais, diminuindo assim, as perdas.

Assim, as técnicas aqui apresentadas e trabalhadas são ferramentas que poderão auxiliar o analista de crédito nas tomadas de decisão, porém nunca serão ferramentas substitutivas, ou seja, por si só não irão dispensar a figura do analista de crédito.

CAPÍTULO VI

6. CONCLUSÕES E SUGESTÕES PARA FUTUROS TRABALHOS

6.1. CONCLUSÕES

As duas técnicas apresentadas e testadas neste trabalho, Árvores de Decisão e Redes Neurais, mostraram ser ferramentas de grande valia para os analistas de crédito bancário. Deste modo, utilizando as informações cadastrais, os analistas têm condições de diagnosticar os novos clientes, quanto ao merecimento de crédito ou não.

Conforme já mencionado no Capítulo II, item 2.1., atualmente o Banco do Brasil analisa a concessão ou não de crédito bancário a pessoas jurídicas através de um aplicativo denominado ANC, o qual não permite a concessão de crédito a empresas que possuam restrições em seu nome ou em nome de seus sócios ou, ainda, no caso em que não forem oferecidas as garantias mínimas exigidas pelos seus normativos internos. Baseado nas informações de cada cliente, o aplicativo ANC sugere valores e garantias para as diversas linhas de crédito. Na presente pesquisa, a questão de valores na concessão de crédito não foi trabalhada.

Os resultados obtidos pelos dois métodos utilizados, comprovaram a sua eficiência na classificação das empresas como adimplentes ou inadimplentes, sendo assim, alcançado o objetivo proposto por este trabalho.

Convém ressaltar que mesmo com uma sinalização totalmente favorável a concessão de crédito a um novo cliente, o mesmo pode vir a se tornar um cliente inadimplente, uma vez que vivemos em um país com uma economia não estabilizada, sujeita a variações. Outros fatores, como por exemplo, um sinistro (incêndio, roubo, ou outro) pode interferir no comportamento da empresa face aos compromissos assumidos.

A tarefa de se conceder ou não crédito, é e sempre será uma tarefa difícil. De qualquer modo, as 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 nas tomadas de decisão, nunca porém poderão por si só, substituir a figura do especialista no contexto da análise de crédito.

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

6.2. SUGESTÕES PARA FUTUROS TRABALHOS

Dentro do contexto trabalhado, os dados das empresas poderiam ser separados por faixas de faturamento anual, a fim de se avaliar melhor o comportamento das empresas de mesmo "tamanho".

As mesmas técnicas apresentadas neste trabalho podem ser utilizadas por exemplo:

- na análise de crédito para pessoas físicas;
- *marketing* direcionado: realizando um trabalho que procure identificar perfil de clientes compradores de determinados produtos bancários;
- ampliar a análise realizada neste trabalho, considerando o valor a ser concedido, assim como o prazo.
- Utilizar os atributos dos nós principais obtidos pela técnica Árvore de Decisão, tendo em vista que são os mais classificadores e formar uma nova base de dados a ser trabalhada com a técnica de Redes Neurais, objetivando uma melhora dos resultados obtidos.

Outras técnicas, como Algoritmo Genético, Análise de Agrupamento, Análise Discriminante Linear de Fisher, assim como a Hibridização de técnicas poderiam ser utilizadas.

Na análise dos resultados obtidos poderia ser utilizada a Lógica Difusa a fim de trabalhar mais detalhadamente os resultados que se encontram na "região de penumbra" (empresas nas fronteiras de classificação).

REFERÊNCIAS BIBLIOGRÁFICAS

1. ADAMOWICZ, E. C. **Reconhecimento de Padrões na Análise Econômico-Financeira de Empresas.** Curitiba, 2000. 110 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia, Concentração em Programação Matemática) - Setor de Tecnologia e Setor de Ciências Exatas, Universidade Federal do Paraná.

2. ADRIAANS, P.; ZANTINGE, D. ***Data Mining.*** England: Addison Wesley Longman, 1996.

3. ALMEIDA, F. C. de. Desvendando o uso de Redes Neurais em problemas de Administração de Empresas. **Revista de Administração de Empresas,** São Paulo, vol. 35, n.1, p. 46-55, jan./fev. 1995.

4. ALMEIDA, F. C. de; SIQUEIRA, J. de O. **Comparação entre regressão logística e redes neurais na previsão de falência de bancos brasileiros.** Terceiro Congresso Brasileiro de Redes Neurais, 4. Florianópolis, p. 1-6, 1997.

5. ALMEIDA, F. C. de; DUMONTIER, P. O uso de Redes Neurais em avaliação de riscos de inadimplência. **Revista de Administração FEA/USP,** vol. 31, n.1, p.52-63, jan./mar. 1996.

6. ANTUNES, C. M. **Árvores de Decisão**, 2002. Disponível em: <http://mega.ist.utl.pt/~ic.apr/doc/aulas/arvoresdecisão.pdf> Acesso em: 21 jul. 2002.

7. AURORA, T. R. P. **Algoritmo de Aprendizado de Máquina**, 2000. Disponível em: <http://inf.ufmg.br/~aurora/tutoriais/arvoresdedecisao.html> Acesso em 18 ago. 2002.

8. BOSIGNOLI, R.; INFANTOSI, A.F.C. Classificação automática do estado de sono ativo neonatal usando Redes Neurais Artificiais. In: XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 1999. Rio de Janeiro, vol. IV, p. 43-45.

9. 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.

10. CARLSON FILHO, C. M.; TAVARES, H. M. F. **Representação de Inequações Lineares com variáveis 0-1 através de Redes Neurais.** XXVII SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 1995.

11. 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.

12. CHEN, M. S.; HAN, J.; YU, P. S. **Data Mining: an overview from database perspective.** IEEE Transaction on Knowledge and Data Engineering, New York, v. 8, n.6, p. 866-883, 1996.

13. CURNOW, G., KOCHMAN, G., MEESTER, S., SARKAR, D., WILTON, K.. **Automating credit and collections decisions at AT&T capital corporation.** Interfaces, v.27, p.29-52, 1997.

14. CUROTTO, C. L. **Uma Estratégia de Data Mining baseada en Indução Incremental de Árvores de Decisão.** COPPE - Programa de Computação de Alto Desempenho. Sistemas Computacionais. Universidade Federal do Rio de Janeiro, 2002.

15. DIN - Departamento de Informática - UEM - Universidade Estadual de Maringá. GSI - Grupo de Sistemas Inteligentes - Mineração de Dados, 1998. Disponível em: <http://www.din.uem.br/ia/mineracao/tecnologia/ferramentas.html> Acesso em: 14 jul. 2002.

16. DW BRASIL **Data Mining.** Disponível em: <http://www.dwbrasil.com.br/dtmining.html> Acesso em: 01 jun. 2002.

17. FAUSETT, L. **Fundamentals of Neural Networks - Architectures, Algorithms, and Applications.** Florida Institute of Technology. Prentice Hall, Upper Saddle River, New Jersey, 1995, 07458.

18. FAYYAD, Usama M., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R.. **Advances in knowledge Discovery & Data Mining.** AAAI/MIT, 1996.

19. FRANCO, R. B.; MARTINS, W. Usando C4.5 e Redes Neurais no acasalamento de Gado Nelore. In: XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 1999. Rio de Janeiro, vol. IV, p. 539-545.

20. FREITAS JÚNIOR, O. de G.; MARTINS, J. G.; RODRIGUES, A. M.; BARCIA, R. M. **Sistema de Apoio à Decisão usando a tecnologia *Data Mining* com estudo de caso da Universidade Estadual de Maringá.** I CONGRESSO BRASILEIRO DE COMPUTAÇÃO - CBCOMP 2001.

21. FREITAS, A. A. Uma Introdução a Data Mining. **Informática Brasileira em Análise.** CESAR - Centro de Estudos e Sistemas Avançados do Recife. Ano II, n. 32, mai./jun. 2000.

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

23. GARCIA, S. C. **O uso de Árvores de Decisão na descoberta de conhecimento na Área da Saúde.** SEMANA ACADÊMICA, 2000. Universidade Federal do Rio Grande do Sul.

24. GORNI, A. A. **Redes Neurais Artificiais - Uma abordagem revolucionária em Inteligência Artificial.** Micro Sistemas, São Paulo, 1993.

25. GUIZZO, Érico. **Quem procura acha.** Revista Negócios Exame. 2000. Disponível em: <http://www.uol.com.br/negociosexame/complementos/revista0002.html> Acesso em: 04 mai. 2002.

26. HADJIPROCOPIS, A. **A Neural Network Implementation on a Transputer System and Applications.** London, London University, 1993. (Dissertação de Mestrado).

27. 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.

28. INMON, W. H. **Como Construir o Data Warehouse.** Rio de Janeiro, Campos, 1997.

29. KRÓSE, B. J. A.; VAN DER SMAGT, P. P. **An Introduction to Neural Networks.** Amsterdam, University of Amsterdam, 1993.

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

31. MENDES FILHO, E. F. **Uma Introdução à Redes Neurais Artificiais.** São Paulo, 1997. Disponível em: <http://sites.uol.com.br/elson_mendes.html> Acesso em: 11 ago. 2002.

32. NIMER, F.; SPANDRI, L.C. **Data Mining.** Revista Developers. Fev./1998, p.32.

33. NOGUEZ, J. H. S. **Um estudo sobre a implementação de Árvore de Decisão com Múltiplos Classificadores.** Porto Alegre, 2000. Trabalho apresentado na Semana Acadêmica 2000, Universidade Federal do Rio Grande do Sul.

34. OSÓRIO, F. **Sistemas Adaptativos Inteligentes - Indução de Árvores de Decisão**, 2000. Disponível em: <http://www.inf.unisinos.br/~osorio/sadi.html> Acesso em: 12 ago. 2002.

35. PLANETA COOPE - Entrevistas. **Entrevista exclusiva com Alessandro Zanasi.** Disponível em: <http://www.planeta.coppe.ufrj.br/entrevistas/entrevista000006.html> Acesso em: 27 jun. 2002.

36. POSSAS, B. A. V.; CARVALHO, M. L. B. de; REZENDE, R. S. F.; MEIRA JR., W. **Data Mining: Técnicas para Exploração de Dados.** Universidade Federal de Minas Gerais, 1998.

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

38. RODRIGUES, A. M. **Escavando Dados no Varejo.** Centro de Estudos em Logística - COPPEAD - Universidade Federal do Rio de Janeiro, 2000.

39. ROSENBERG, E. & GLEIT, A. **Quantitative methods in credit management: a survey.** Operations Research, v.42, n.4, p.589-613, 1994.

40. SILVA, E. M. **Avaliação do Estado da Arte e Produtos Data Mining.** UCB - Universidade Católica de Brasília, 2000.

41. SIMEÃO, J. de M. **Bancos de Dados Geográficos e Redes Neurais Artificiais: Tecnologias de Apoio à Gestão de Território.** São Paulo, 1999. Universidade de São Paulo.

42. STEINER, M. T. A. **Redes Neurais.** Universidade Federal do Paraná. Métodos Numéricos em Engenharia - Pesquisa Operacional, 1999.

43. 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**, São Paulo, vol. 34, n.3, p. 56-67, jul./set. 1999.

44. TATIBANA, C. Y. ; KAETSY, D. Y. **Uma Introdução às Redes Neurais.** 2000. Disponível em: <http://www.din.uem.br/ia/neurais/#artificial.html> Acesso em: 14 jul. 2002.

45. TAFNER, M. A. **Redes Neurais Artificiais: Aprendizado e Plasticidade.** Revista Cérebro & Mente, 2(5), mar./mai. 1998.

46. TERRA, A. R. T.; PEREIRA, N. A. **Programação da Produção: uma abordagem por Redes Neurais Artificiais.** São Carlos (SP), 1999.

47. THEARLING, Kurt, BERSON, Alex, SMITH, Stephen. **Building Data Mining Appllications for CRM.** McGraw Hill, 2000.

48. WITTEN, I. H.; FRANK, E. **Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations.** Morgan Kaufmann Publishers. San Francisco, California, 2000.

ANEXO I

FORMULÁRIO UTILIZADO PARA COLETA DOS DADOS

EMPRESA: _____

RESTRIÇÕES: ( ) SIM ( ) NÃO

RESTRIÇÕES BAIXADAS NOS 5 ÚLTIMOS ANOS: ( ) SIM ( ) NÃO

TEMPO DE CONTA: _____anos e _____ meses

SETOR DE ATIVIDADE: _____TEMPO ATIVIDADE: _____

NÚMERO DE FUNCIONÁRIOS: _____

SEDE DA EMPRESA: ( ) PRÓPRIA ( ) ALUGADA ( ) CEDIDA

BAIRRO: _____

PRINCIPAIS CLIENTES: ( ) pessoas físicas ( ) pessoas jurídicas ( ) misto

CONCENTRAÇÃO VENDAS 5 PRINCIPAIS CLIENTES: ( ) menos de 20%

( ) mais de 20%

FATURAMENTO BRUTO ANUAL: R$ _____

CLIENTE EM OUTRO BANCO: ( ) SIM ( ) NÃO

BENS IMÓVEIS: R$ _____BENS MÓVEIS: R$ _____

CRÉDITO JUNTO A FORNECEDORES: ( ) SIM ( ) NÃO

SEGURO EMPRESARIAL: ( ) SIM ( ) NÃO

APLICAÇÕES FINANCEIRAS NO BB: ( ) SIM ( ) NÃO

VENDAS A PRAZO: ( ) menos de 20% ( ) mais de 20%

EXPERIÊNCIA DE CRÉDITO NO BB: ( ) SIM ( ) NÃO

HISTÓRICO DA CONTA CORRENTE: _____

CONCEITO NA PRAÇA: _____

PONTUALIDADE: _____

SÓCIOS C/RESTRIÇÕES: ( ) SIM ( ) NÃO

SÓCIOS C/RESTRIÇÕES 5 ÚLTIMOS ANOS: ( ) SIM ( ) NÃO

SOCIEDADE ENTRE CÔNJUGES: ( ) SIM ( ) NÃO

BENS DOS SÓCIOS: IMÓVEIS: R$ _____MÓVEIS: R$ _____

RISCO: _____ ( ) ADIMPLENTE ( ) INADIMPLENTE

ANEXO II

TABELA A1 - DADOS CADASTRAIS DAS 339 EMPRESAS

Empre-sas	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	2	2	17	1	5	5	2	1	1	12.277,00	2	0,00	0,00	2	5	2	1	2	2	1	1	0,00	0,00	5	1
2	2	2	76	1	4	1	2	1	1	56.847,00	2	0,00	0,00	2	5	2	2	1	2	1	1	407.000,00	0,00	2	1
3	2	2	175	2	5	8	1	2	2	157.855,00	2	115.750,00	0,00	2	5	2	2	1	2	2	1	0,00	0,00	1	1
4	2	2	51	1	4	43	1	2	1	583.170,00	1	0,00	0,00	1	5	2	1	1	2	2	2	92.800,00	167.500,00	1	1
5	2	2	13	3	5	0	1	1	1	292.591,00	1	0,00	0,00	1	5	2	3	1	2	2	2	20.420,00	0,00	2	1
6	2	2	6	1	5	0	2	1	1	113.350,00	2	0,00	0,00	2	5	2	3	1	2	2	2	12.000,00	0,00	3	1
7	2	2	0	3	1	0	3	2	1	49.308,00	2	0,00	0,00	1	5	1	3	3	2	2	2	0,00	0,00	3	1
8	2	2	28	1	3	0	2	1	1	98.464,00	1	0,00	0,00	2	5	1	3	1	2	2	2	11.500,00	6.500,00	3	1
9	2	1	27	1	5	2	2	1	1	156.150,00	2	0,00	0,00	2	5	2	2	1	2	1	2	36.500,00	0,00	2	1
10	2	2	17	1	2	12	2	1	1	10.653.948,00	1	0,00	0,00	1	5	2	1	1	2	2	1	555.820,00	95.000,00	1	1
11	2	2	13	1	2	25	2	2	1	4.128.894,00	1	0,00	0,00	1	5	2	1	1	2	1	2	1.783.520,00	167.000,00	1	1
12	2	2	165	1	5	21	3	2	1	2.661.633,00	1	0,00	0,00	1	5	2	2	1	2	2	2	8.906.093,73	74.000,00	1	1
13	2	2	22	2	1	3	1	1	1	78.643,00	1	0,00	28.000,00	1	5	2	1	1	2	2	1	40.000,00	16.800,00	2	1
14	2	2	0	2	2	14	3	2	2	222.600,00	1	0,00	0,00	2	5	2	3	3	2	2	2	160.000,00	51.500,00	2	1
15	1	2	135	3	5	8	3	2	1	123.467,00	2	0,00	0,00	1	5	2	2	1	2	2	2	0,00	0,00	1	1
16	2	2	50	1	3	1	2	1	1	119.915,00	1	0,00	0,00	2	5	2	2	1	2	2	1	5.000,00	32.850,00	1	1
17	2	2	39	1	3	1	2	1	1	55.537,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
18	2	2	172	1	5	17	1	1	1	250.000,00	1	720.951,00	0,00	1	5	2	2	1	1	2	2	132.000,00	222.500,00	1	1
19	2	2	12	1	2	0	3	2	1	37.032,00	2	0,00	0,00	1	1	2	1	1	2	2	2	0,00	0,00	2	1
20	2	2	54	1	3	1	3	1	1	89.635,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
21	2	2	56	1	4	0	3	2	2	41.227,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
22	2	2	3	1	1	0	2	2	1	450.000,00	1	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	4	1
23	2	2	0	1	3	0	3	2	1	119.927,00	1	0,00	0,00	2	1	1	3	3	2	1	1	17.208,90	0,00	2	1
24	2	2	40	3	5	5	3	2	1	330.897,00	1	0,00	25.000,00	1	5	1	2	1	2	2	1	0,00	0,00	1	1
25	2	2	73	3	4	6	2	1	1	86.239,00	1	0,00	7.400,00	2	5	1	3	1	2	2	1	0,00	0,00	3	1
26	2	2	12	2	4	6	2	1	1	27.431,00	2	0,00	10.738,00	2	5	2	2	1	1	2	1	0,00	0,00	3	1
27	2	2	30	2	5	4	3	2	2	47.729,00	2	0,00	0,00	1	5	2	2	1	2	2	2	0,00	0,00	1	1
28	2	2	48	3	5	0	2	1	2	18.146,00	2	0,00	0,00	2	5	1	3	1	2	2	2	0,00	0,00	2	1
29	2	2	41	1	3	1	2	2	1	110.195,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1

30	2	2	0	3	1	0	3	2	1	27.787,00	2	0,00	0,00	2	5	2	3	3	2	2	1	0,00	0,00	3	1
31	2	1	40	3	5	11	3	2	1	284.100,00	2	0,00	15.000,00	1	5	2	2	1	2	1	1	0,00	13.000,00	1	1
32	2	2	95	1	5	4	3	2	1	1.117.908,00	2	0,00	0,00	2	5	2	2	1	2	2	1	122.394,00	0,00	1	1
33	2	2	22	1	3	6	3	1	1	2.325.996,00	1	0,00	0,00	2	4	2	3	1	2	1	1	0,00	0,00	2	1
34	2	1	57	1	4	6	2	1	1	1.544.723,00	1	0,00	0,00	2	5	2	2	1	2	1	1	0,00	0,00	2	1
35	2	2	48	1	4	11	3	2	1	190.989,00	2	0,00	0,00	2	1	2	2	1	2	2	1	140.000,00	51.700,00	1	1
36	2	1	13	2	3	9	2	2	1	863.204,00	2	0,00	0,00	1	5	2	1	1	2	1	2	15.000,00	0,00	2	1
37	2	2	161	3	5	4	2	1	1	108.619,00	1	5.087,00	0,00	2	5	2	3	1	2	2	1	8.000,00	28.486,00	1	1
38	2	2	59	1	3	1	2	1	1	52.726,00	2	0,00	0,00	2	1	2	3	1	2	2	1	79.002,43	49.650,00	1	1
39	2	2	50	3	3	6	2	1	1	116.231,00	2	0,00	0,00	2	1	2	2	1	2	2	2	0,00	0,00	3	1
40	2	2	55	3	5	11	1	1	1	45.763,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	1	1
41	2	2	78	3	5	20	2	2	1	804.093,00	2	0,00	0,00	2	5	2	2	1	2	1	2	167.500,00	0,00	1	1
42	2	2	40	1	3	10	2	2	1	160.023,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	3	1
43	2	2	51	1	4	18	3	2	1	1.539.738,00	1	0,00	0,00	2	5	2	2	1	2	2	2	80.000,00	9.000,00	2	1
44	1	2	42	2	5	26	2	1	1	1.106.427,00	1	0,00	0,00	2	4	1	3	1	1	1	2	399.900,00	0,00	2	1
45	2	2	62	1	5	2	2	1	1	213.007,00	2	0,00	0,00	1	5	2	2	1	2	2	1	0,00	66.800,00	1	1
46	2	2	0	1	5	0	3	1	1	25.339,00	2	0,00	0,00	2	1	2	3	3	2	2	1	0,00	0,00	3	1
47	2	1	168	1	5	15	2	1	1	338.317,00	1	0,00	0,00	2	5	2	2	1	2	2	1	0,00	56.500,00	1	1
48	2	2	16	1	2	0	1	2	2	58.977,00	2	0,00	0,00	2	5	2	3	1	2	2	1	15.000,00	1.500,00	2	1
49	2	2	29	1	3	1	3	2	1	144.744,00	1	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
50	2	2	316	1	5	25	3	1	1	1.074.992,00	1	0,00	0,00	2	4	2	3	1	1	1	2	0,00	0,00	1	1
51	2	2	0	2	1	5	3	2	2	108.000,00	2	0,00	0,00	2	5	2	3	3	2	2	1	0,00	4.000,00	3	1
52	2	2	0	1	1	1	2	2	1	48.944,00	2	0,00	0,00	1	5	2	3	3	2	2	2	0,00	0,00	2	1
53	2	2	29	2	5	3	2	2	1	833.974,00	1	10.359,00	32.000,00	1	5	2	2	1	2	2	2	0,00	0,00	1	1
54	2	2	132	2	5	25	3	2	1	1.393.590,00	1	0,00	0,00	2	5	2	2	1	2	1	2	164.000,00	0,00	2	1
55	2	2	3	1	3	5	2	1	1	203.761,00	1	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	2	1
56	2	2	56	1	3	4	2	1	1	460.586,00	1	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	2	1
57	2	2	0	3	3	2	1	1	1	52.573,00	1	60.000,00	0,00	2	5	2	3	3	2	2	1	20.000,00	0,00	1	1
58	2	2	13	1	2	1	2	2	1	151.682,00	2	0,00	0,00	1	5	2	1	4	2	2	1	42.000,00	8.000,00	3	1
59	2	2	22	3	2	0	2	1	1	107.815,00	2	0,00	0,00	2	5	1	3	1	1	1	2	0,00	0,00	3	1
60	2	2	19	3	1	0	3	2	1	1.178.772,00	2	98.000,00	0,00	2	5	1	3	1	2	2	2	180.000,00	0,00	2	1
61	2	2	33	1	4	3	3	2	1	249.171,00	2	0,00	11.113,00	1	1	2	2	1	2	2	2	0,00	0,00	2	1
62	1	2	46	3	3	10	2	1	1	1.309.773,00	2	0,00	0,00	2	5	2	2	1	2	2	2	30.000,00	0,00	2	1
63	2	2	175	3	5	5	2	2	1	244.613,00	1	0,00	0,00	2	5	2	3	1	2	2	2	0,00	22.000,00	1	1
64	2	2	40	3	3	8	3	1	1	44.553,00	1	0,00	0,00	2	4	2	1	1	2	1	2	300.000,00	0,00	2	1
65	2	2	24	3	4	10	2	2	1	242.866,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	3	1
66	2	2	0	2	1	0	2	2	2	149.093,00	2	0,00	0,00	2	5	2	3	3	2	2	2	0,00	15.000,00	3	1

67	2	2	201	1	5	22	3	2	1	3.430.892,00	1	75.000,00	0,00	2	4	1	2	1	2	2	1	0,00	0,00	1	1
68	2	2	138	2	5	8	1	1	1	459.663,00	1	0,00	0,00	1	1	2	3	1	2	2	1	380.862,00	41.000,00	1	1
69	2	2	57	1	5	0	2	1	1	44.495,00	2	0,00	0,00	2	5	2	3	1	2	2	1	37.564,00	18.900,00	2	1
70	2	2	317	3	5	0	3	1	2	104.819,00	1	120.000,00	0,00	2	5	1	3	1	2	2	1	50.000,00	0,00	1	1
71	2	2	13	1	2	0	3	2	1	438.502,00	2	0,00	0,00	2	5	2	1	1	2	2	1	0,00	0,00	3	1
72	2	2	0	1	3	11	3	1	1	53.489,00	1	0,00	0,00	2	5	2	3	3	2	2	1	0,00	14.000,00	3	1
73	2	2	0	1	5	1	2	1	1	12.228,00	2	0,00	0,00	2	5	1	3	3	2	1	2	0,00	0,00	2	1
74	2	2	15	2	5	1	2	1	1	138.069,00	2	38.000,00	0,00	2	5	2	3	1	2	2	2	40.000,00	37.000,00	1	1
75	2	2	26	2	3	6	1	2	1	1.124.383,00	2	0,00	0,00	2	5	2	2	1	2	2	2	22.700,00	49.680,00	1	1
76	2	2	34	2	3	7	3	2	1	176.884,00	1	0,00	17.200,00	1	5	2	2	1	2	2	1	207.028,00	30.000,00	1	1
77	2	2	0	1	1	0	3	2	1	192.000,00	2	0,00	0,00	2	1	1	3	3	2	2	2	18.200,00	0,00	2	1
78	2	2	0	3	1	1	3	2	1	35.450,00	2	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	3	1
79	2	2	315	1	5	30	1	1	1	6.173.748,00	1	760.000,00	0,00	2	5	2	3	1	2	1	2	185.536,00	257.000,00	1	1
80	2	2	37	2	3	22	3	2	1	140.360,00	2	0,00	40.000,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
81	2	2	55	3	4	2	2	1	1	95.325,00	1	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	1	1
82	2	2	56	3	5	20	1	2	1	450.302,00	1	183.100,00	0,00	1	5	2	2	1	2	2	1	5.000,00	0,00	1	1
83	2	2	40	1	4	1	3	1	1	59.357,00	2	0,00	0,00	2	5	1	1	1	2	2	2	0,00	0,00	2	1
84	2	2	92	1	5	10	1	1	1	1.552.039,00	1	0,00	0,00	2	5	2	2	1	2	2	1	0,00	0,00	1	1
85	2	2	55	1	3	7	3	1	1	348.456,00	1	0,00	10.500,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
86	2	2	144	1	5	1	2	1	1	58.349,00	2	0,00	0,00	2	5	2	2	1	2	2	1	15.000,00	0,00	2	1
87	1	2	59	1	4	0	3	1	1	162.966,00	2	0,00	0,00	2	5	2	2	1	2	1	1	43.370,00	0,00	3	1
88	2	2	11	1	4	0	2	1	1	18.226,00	2	0,00	0,00	2	5	2	1	1	2	2	2	0,00	0,00	3	1
89	2	2	296	2	5	5	3	2	1	76.099,00	1	0,00	0,00	1	1	2	2	1	2	2	1	77.000,00	0,00	1	1
90	2	2	48	1	3	4	3	1	1	1.300.598,00	2	0,00	0,00	2	5	2	2	1	2	2	1	122.394,00	0,00	1	1
91	2	2	17	1	4	4	3	2	1	1.488.261,00	2	0,00	0,00	2	5	2	1	1	2	2	1	122.394,00	0,00	1	1
92	2	2	33	1	3	11	2	2	1	603.604,00	1	0,00	90.750,00	2	5	2	2	1	2	2	2	7.500,00	0,00	2	1
93	2	2	30	1	3	6	2	2	1	427.600,00	2	0,00	68.000,00	1	5	2	2	4	2	2	1	0,00	16.000,00	2	1
94	2	2	27	1	3	3	2	2	2	770.955,00	2	0,00	0,00	2	5	2	2	1	2	2	0	0,00	11.200,00	2	1
95	2	2	37	3	3	2	3	1	1	38.110,00	2	0,00	0,00	2	5	2	2	1	2	2	1	0,00	0,00	2	1
96	2	2	315	1	5	4	1	1	1	143.461,00	2	0,00	0,00	1	5	1	2	1	2	2	2	0,00	0,00	1	1
97	2	2	70	2	4	60	3	2	1	1.084.320,00	1	0,00	0,00	2	5	2	2	1	2	2	1	0,00	9.600,00	1	1
98	2	2	300	3	5	0	3	1	1	300.601,00	1	0,00	0,00	1	5	2	2	1	2	2	2	0,00	173.615,23	1	1
99	2	2	25	1	3	4	2	1	1	464.321,00	2	0,00	48.000,00	1	5	2	2	1	2	2	1	0,00	0,00	1	1
100	2	2	48	1	5	3	2	1	1	146.531,00	2	0,00	0,00	1	5	2	2	1	2	2	2	0,00	0,00	1	1
101	2	2	1	1	1	1	2	2	1	148.843,00	2	0,00	0,00	1	1	2	3	3	2	2	2	0,00	0,00	3	1
102	2	2	10	1	2	0	2	2	1	36.758,00	2	0,00	0,00	2	5	2	3	1	2	2	2	0,00	151.800,00	2	1
103	2	2	43	3	3	1	1	1	1	97.271,00	2	0,00	0,00	2	5	2	2	1	2	1	1	40.000,00	14.600,00	2	1

104	2	2	43	1	5	55	1	1	1	3.323.480,00	1	640.600,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	1	1
105	2	2	288	1	5	3	3	1	1	35.472,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
106	2	2	42	1	4	1	2	2	1	259.116,00	2	7.200,00	0,00	1	5	2	2	1	2	2	2	0,00	0,00	1	1
107	2	2	0	1	5	0	2	1	1	103.796,00	1	0,00	0,00	1	5	2	1	3	2	2	2	0,00	0,00	2	1
108	2	2	66	3	4	0	2	1	1	127.919,00	2	0,00	0,00	1	5	2	2	1	2	2	2	12.700,00	0,00	1	1
109	2	2	239	3	5	23	1	2	1	268.587,00	1	210.000,00	0,00	1	5	2	2	1	2	2	2	0,00	0,00	1	1
110	2	2	13	3	5	4	1	2	1	147.579,00	1	0,00	0,00	2	5	2	1	1	2	2	1	8.286,00	44.000,00	1	1
111	2	2	92	1	4	15	3	2	1	1.403.611,00	1	0,00	0,00	1	5	2	2	1	2	1	2	138.700,00	0,00	1	1
112	2	2	13	1	3	0	1	2	1	119.980,00	2	0,00	0,00	2	4	2	1	1	2	2	2	4.800,00	128.600,00	1	1
113	2	2	49	3	3	16	3	2	1	296.360,00	2	0,00	0,00	1	5	2	2	1	2	2	1	0,00	27.000,00	1	1
114	2	2	22	1	2	1	2	1	1	10.864,00	2	0,00	0,00	1	5	2	1	1	2	2	2	0,00	9.500,00	3	1
115	2	2	65	2	5	26	1	2	1	447.790,00	2	108.000,00	0,00	1	5	2	2	1	2	1	2	0,00	0,00	2	1
116	2	2	5	1	1	4	2	1	1	73.077,00	1	0,00	0,00	2	5	1	3	1	2	2	2	0,00	10.000,00	2	1
117	1	2	50	3	5	4	3	2	1	35.158,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
118	2	2	31	3	3	5	2	1	1	50.917,00	1	0,00	0,00	2	5	2	2	1	2	2	2	0,00	12.850,00	2	1
119	2	2	4	1	1	10	3	2	1	405.399,00	2	0,00	0,00	2	5	2	3	1	2	2	2	171.264,00	67.200,00	2	1
120	2	2	24	1	5	1	3	2	1	82.076,00	2	63.185,00	0,00	2	5	2	2	1	2	1	1	19.200,00	7.000,00	2	1
121	2	2	80	2	5	15	3	1	1	670.873,00	2	0,00	0,00	1	2	2	2	1	2	2	1	598.800,00	0,00	1	1
122	2	2	0	3	3	1	2	1	1	19.605,00	1	0,00	0,00	2	5	1	3	3	2	2	2	0,00	0,00	3	1
123	2	2	34	1	5	7	3	2	1	225.897,00	2	0,00	0,00	2	5	2	2	1	2	1	1	75.000,00	8.000,00	1	1
124	2	2	98	3	5	0	1	2	1	96.490,00	1	80.520,00	0,00	1	5	1	3	1	2	2	1	0,00	0,00	1	1
125	2	1	124	2	5	6	3	2	1	81.584,00	1	90.000,00	0,00	2	5	2	2	1	2	2	2	0,00	16.000,00	2	1
126	2	2	34	2	5	9	1	1	1	161.018,00	1	39.000,00	9.500,00	2	5	2	2	1	2	2	1	0,00	0,00	1	1
127	2	2	30	1	5	1	2	2	1	117.625,00	2	0,00	13.500,00	1	5	2	2	1	2	2	1	0,00	11.500,00	1	1
128	2	2	3	1	1	0	2	1	1	38.177,00	1	0,00	18.000,00	1	1	1	3	1	2	2	2	0,00	0,00	3	1
129	2	2	0	1	5	6	1	1	1	140.870,00	1	280.000,00	0,00	2	5	1	3	3	2	2	1	0,00	0,00	2	1
130	2	2	61	1	3	18	2	2	1	502.270,00	1	0,00	0,00	1	5	2	2	1	2	2	2	42.600,00	0,00	2	1
131	2	2	40	1	3	1	3	2	1	41.154,00	1	0,00	0,00	1	5	2	2	4	2	2	2	0,00	0,00	1	1
132	2	2	1	2	2	5	2	2	2	47.242,00	2	0,00	0,00	1	5	2	3	3	2	2	2	0,00	15.000,00	4	1
133	2	2	1	1	1	0	2	1	1	443.000,00	1	0,00	0,00	2	4	2	3	3	2	2	1	0,00	80.250,00	2	1
134	2	2	53	1	3	3	3	1	1	70.592,00	2	0,00	0,00	1	5	2	2	1	2	1	2	130.020,00	0,00	2	1
135	2	2	7	2	1	5	3	2	1	322.000,00	2	0,00	25.500,00	1	5	2	1	1	2	2	2	0,00	7.500,00	3	1
136	2	2	0	3	1	0	3	2	1	50.858,00	2	0,00	0,00	2	5	1	3	3	2	2	1	0,00	0,00	3	1
137	2	2	0	1	3	0	3	2	1	51.719,00	1	0,00	0,00	2	5	2	3	3	2	2	1	0,00	10.750,00	3	1
138	2	2	2	3	1	1	2	2	1	25.765,00	2	0,00	0,00	1	5	1	3	3	2	2	2	45.415,00	0,00	3	1
139	2	2	188	3	5	26	1	1	1	427.082,00	2	320.000,00	0,00	2	5	2	2	1	2	2	1	41.670,00	0,00	1	1
140	2	2	101	3	4	17	3	2	1	244.676,00	2	0,00	0,00	2	5	2	2	1	2	2	2	352.456,14	0,00	1	1

141	2	2	27	3	4	32	3	2	1	192.949,00	1	0,00	0,00	1	5	2	2	1	2	1	1	130.020,00	0,00	1	1
142	2	1	31	2	5	15	3	1	1	414.974,00	2	0,00	112.600,00	2	5	2	2	1	2	1	1	213.043,00	0,00	2	1
143	2	2	88	2	5	0	3	2	1	1.017.061,00	1	0,00	0,00	2	5	2	3	1	2	2	2	366.088,00	24.000,00	2	1
144	2	2	12	2	5	28	1	2	1	488.747,00	2	52.000,00	0,00	1	5	2	1	1	2	2	1	0,00	6.700,00	1	1
145	1	2	121	2	5	27	3	1	1	280.151,00	2	0,00	0,00	2	5	1	3	1	2	2	2	0,00	0,00	3	1
146	2	1	24	1	3	2	1	1	2	463.050,00	2	0,00	10.350,00	2	5	1	1	1	2	1	2	0,00	0,00	3	1
147	2	2	34	2	4	14	3	2	2	266.135,00	2	0,00	0,00	2	5	1	1	1	2	1	1	0,00	0,00	2	1
148	2	2	323	1	5	0	1	1	1	261.589,00	1	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	2	1
149	2	2	47	1	3	5	2	1	1	126.786,00	2	100.000,00	0,00	2	5	1	3	1	2	2	1	0,00	0,00	1	1
150	2	2	17	1	3	3	2	2	1	216.021,00	2	0,00	0,00	2	4	2	1	1	2	2	1	0,00	0,00	3	1
151	2	2	73	3	4	4	3	2	1	26.523,00	1	0,00	159.000,00	1	5	1	2	1	2	2	2	40.000,00	92.000,00	2	1
152	2	2	150	1	5	16	1	1	1	277.711,00	1	315.000,00	0,00	2	5	2	2	1	2	1	2	378.000,00	174.600,00	1	1
153	1	2	144	1	5	0	2	1	1	119.838,00	2	0,00	0,00	2	5	1	3	1	2	2	1	0,00	25.228,00	2	1
154	2	2	12	3	4	4	3	1	1	22.853,00	2	0,00	11.970,00	2	5	1	1	1	2	2	2	0,00	0,00	2	1
155	2	2	0	1	1	0	2	1	1	66.176,00	2	0,00	0,00	2	5	1	3	3	2	2	2	0,00	0,00	3	1
156	2	2	0	1	2	0	2	1	1	113.500,00	2	0,00	0,00	2	5	1	3	3	2	1	1	0,00	0,00	5	1
157	2	2	0	2	5	0	3	2	2	56.214,00	1	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	3	1
158	2	2	11	1	3	1	2	1	1	22.254,00	1	0,00	0,00	1	5	1	3	1	2	2	2	0,00	0,00	3	1
159	2	2	0	3	2	0	2	2	1	221.105,00	2	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	4	1
160	2	2	310	3	5	19	1	2	1	233.428,00	1	0,00	0,00	1	1	2	2	1	2	1	2	0,00	0,00	2	1
161	2	2	44	2	3	6	3	2	1	33.849,00	2	0,00	3.000,00	1	5	2	2	1	2	2	2	0,00	0,00	1	1
162	2	2	45	3	4	3	2	1	1	94.112,00	1	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	1	1
163	2	2	0	2	5	33	3	2	1	784.027,00	1	0,00	11.000,00	2	5	2	3	3	2	2	2	1.526.000,00	13.000,00	1	1
164	2	2	25	3	3	28	2	1	1	401.550,00	1	0,00	764.550,00	1	5	2	2	1	2	1	2	0,00	0,00	2	1
165	2	2	36	1	3	2	2	2	1	94.168,00	2	0,00	0,00	2	5	2	2	2	2	2	1	17.220,00	0,00	2	1
166	2	2	15	1	5	0	3	1	1	1.685.464,00	1	0,00	0,00	2	4	2	1	1	2	1	2	0,00	0,00	2	1
167	2	2	232	3	5	3	3	1	1	42.335,00	2	0,00	26.100,00	1	4	1	2	1	2	2	1	0,00	0,00	2	1
168	2	2	1	1	1	0	3	2	1	38.723,00	1	0,00	21.300,00	1	5	2	1	3	2	2	2	0,00	0,00	4	1
169	2	2	41	1	5	0	3	2	1	113.032,00	1	0,00	0,00	2	2	1	3	1	2	2	1	0,00	0,00	2	1
170	2	2	107	5	1	0	3	2	1	270.680,00	1	0,00	0,00	2	5	2	2	1	2	2	1	34.100,00	0,00	2	1
171	2	2	15	1	3	2	2	1	1	242.716,00	1	0,00	0,00	1	2	2	1	1	2	2	2	0,00	0,00	2	1
172	2	2	12	1	3	7	2	1	1	46.445,00	1	0,00	0,00	2	5	1	1	1	2	2	2	0,00	0,00	3	1
173	2	2	1	1	1	2	2	1	1	59.988,00	1	0,00	0,00	2	5	1	3	3	2	2	2	0,00	0,00	3	1
174	1	2	82	1	4	0	2	1	1	90.327,00	1	0,00	0,00	1	5	2	3	1	2	2	2	0,00	0,00	3	1
175	2	2	318	1	5	13	3	1	1	778.294,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
176	2	2	25	1	5	12	2	2	1	806.198,00	1	40.000,00	0,00	1	5	2	2	1	2	2	2	72.000,00	11.600,00	1	1
177	2	2	6	3	1	0	3	2	1	94.948,00	2	0,00	0,00	1	5	2	3	1	2	2	2	0,00	12.500,00	3	1
178	2	2	0	3	1	0	2	2	1	21.509,00	2	0,00	0,00	2	1	2	3	3	2	2	1	0,00	0,00	3	1

179	2	2	7	2	1	1	2	1	2	51.730,00	2	0,00	0,00	2	5	2	3	1	2	2	2	0,00	2.000,00	4	1
180	2	2	14	3	3	2	3	2	2	53.620,00	2	0,00	0,00	1	5	2	3	1	2	2	2	0,00	0,00	3	1
181	1	2	103	1	5	0	3	2	1	76.241,00	2	0,00	7.200,00	2	1	2	3	1	2	2	1	12.100,00	0,00	3	1
182	2	2	166	1	5	2	3	1	1	26.455,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	1	1
183	2	2	2	1	1	0	2	1	1	71.303,00	2	0,00	0,00	2	5	1	3	1	2	2	2	0,00	0,00	4	1
184	2	2	0	1	2	3	2	1	1	84.673,00	2	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	3	1
185	2	2	15	3	5	0	3	1	1	185.880,00	1	0,00	64.000,00	1	5	1	1	1	2	2	2	0,00	0,00	2	1
186	2	2	86	1	5	0	1	1	1	14.054,00	2	0,00	0,00	2	1	1	3	1	1	1	2	0,00	0,00	2	1
187	2	2	40	2	5	7	3	2	1	211.364,00	1	54.032,00	0,00	2	5	2	2	1	2	2	2	0,00	13.700,00	2	1
188	2	2	147	1	5	11	3	2	1	223.283,00	2	0,00	0,00	2	5	2	2	1	2	2	1	30.354,00	0,00	2	1
189	2	2	40	3	5	14	3	2	1	162.674,00	1	12.880,00	0,00	2	5	2	2	1	2	1	2	11.720,00	17.400,00	2	1
190	1	2	33	3	3	16	3	2	1	454.265,00	2	0,00	0,00	1	5	2	2	1	2	2	2	4.500,00	18.000,00	2	1
191	1	2	39	1	3	6	2	2	1	73.206,00	2	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	3	1
192	2	2	12	2	3	6	2	2	1	149.381,00	2	0,00	4.000,00	1	5	2	1	1	2	2	1	4.000,00	10.800,00	2	1
193	2	2	98	2	5	17	1	2	1	451.487,00	1	115.000,00	19.000,00	1	5	2	2	1	2	2	2	0,00	6.000,00	1	1
194	2	2	13	1	5	0	3	2	1	216.510,00	1	0,00	0,00	2	3	2	1	1	2	2	2	0,00	0,00	2	1
195	2	2	87	2	1	8	1	2	2	124.452,00	1	0,00	0,00	2	1	2	2	1	2	2	2	10.000,00	0,00	2	1
196	1	2	34	2	5	15	1	2	2	505.681,00	2	198.970,00	0,00	2	5	2	2	1	2	2	2	0,00	0,00	2	1
197	2	2	49	3	5	3	2	1	1	106.569,00	1	0,00	31.355,00	2	5	2	2	1	2	2	2	0,00	0,00	1	1
198	2	2	68	2	4	3	2	2	1	44.139,00	2	0,00	0,00	2	1	2	2	1	2	2	2	0,00	0,00	2	1
199	2	2	0	3	1	0	3	2	1	35.000,00	1	0,00	7.314,00	2	5	1	3	3	2	2	2	0,00	0,00	4	1
200	2	2	24	2	5	3	3	2	1	56.541,00	1	67.700,00	9.000,00	1	5	2	2	1	2	2	1	0,00	0,00	2	1
201	2	2	0	3	3	0	3	2	1	48.000,00	2	0,00	0,00	1	5	2	3	3	1	2	2	0,00	5.500,00	5	1
202	2	2	6	3	2	1	3	2	1	36.970,00	2	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	3	1
203	2	2	55	1	5	0	2	1	1	327.314,00	1	0,00	0,00	2	1	2	3	1	2	2	2	0,00	0,00	1	1
204	2	1	73	1	4	16	3	2	1	479.953,00	1	0,00	0,00	2	5	2	2	4	2	2	2	0,00	0,00	3	1
205	2	2	0	1	4	0	2	2	1	6.082,00	1	0,00	0,00	2	5	2	3	3	2	1	2	0,00	0,00	4	1
206	2	2	29	1	3	1	2	2	1	64.113,00	2	0,00	0,00	2	5	1	2	1	2	2	2	0,00	15.500,00	2	1
207	2	2	24	1	3	2	1	2	1	401.522,00	2	80.000,00	0,00	2	5	2	2	1	2	2	1	0,00	7.500,00	1	1
208	2	2	46	1	3	1	2	1	1	98.445,00	1	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	2	1
209	2	2	0	3	1	0	2	1	1	18.992,00	2	0,00	0,00	2	5	1	3	3	2	2	2	0,00	0,00	3	1
210	2	2	12	1	2	0	2	1	1	31.818,00	1	0,00	0,00	2	5	2	1	1	2	2	2	0,00	0,00	3	1
211	2	2	6	1	1	2	3	1	1	178.787,00	1	0,00	0,00	2	5	2	1	1	2	2	2	0,00	0,00	2	1
212	2	2	1	2	1	0	1	2	1	223.275,00	2	0,00	0,00	2	1	2	1	1	2	2	2	0,00	0,00	3	1
213	2	2	0	3	1	0	3	1	2	55.287,00	1	0,00	0,00	2	1	2	3	3	2	2	2	0,00	58.500,00	4	1
214	2	2	28	2	5	2	3	2	1	11.000,00	2	0,00	0,00	1	5	1	2	1	2	2	2	0,00	0,00	2	1
215	1	2	31	2	3	4	3	2	1	714.642,00	2	0,00	10.324,00	1	5	2	2	1	2	1	2	120.809,47	22.600,00	2	1
216	2	2	9	1	3	0	2	1	1	731.160,00	1	0,00	0,00	1	5	2	3	1	2	2	2	105.260,00	0,00	1	1

217	2	2	311	2	5	12	3	2	1	884.783,00	1	70.000,00	0,00	1	5	2	2	1	2	2	1	120.809,47	0,00	1	1
218	2	2	311	2	5	12	3	2	1	400.623,00	1	0,00	0,00	1	1	2	2	1	2	2	2	0,00	0,00	1	1
219	2	2	0	1	1	0	3	2	1	97.308,00	2	0,00	0,00	2	5	1	3	3	2	2	2	0,00	0,00	4	1
220	2	2	0	3	1	0	3	2	1	30.818,00	1	0,00	0,00	2	5	1	3	3	2	2	2	0,00	0,00	4	1
221	2	2	40	3	5	23	3	2	1	251.919,00	2	0,00	0,00	2	5	2	2	1	2	2	1	34.033,00	0,00	1	1
222	2	2	4	1	1	0	2	2	2	48.400,00	1	0,00	0,00	2	5	2	3	1	2	1	1	0,00	28.000,00	2	1
223	2	2	13	1	2	2	3	2	2	100.710,00	1	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	3	1
224	2	2	14	3	2	9	2	2	1	85.350,00	2	0,00	0,00	2	5	2	1	1	2	1	2	5.500,00	0,00	3	1
225	2	2	33	3	3	0	2	2	1	691.101,00	2	0,00	0,00	2	4	2	3	1	2	2	2	0,00	0,00	3	1
226	2	2	14	2	2	8	2	2	1	247.693,00	2	0,00	0,00	1	5	2	1	1	2	2	2	0,00	0,00	4	1
227	2	2	0	2	1	4	1	2	1	483.000,00	1	0,00	22.000,00	1	5	2	3	3	2	2	2	0,00	30.000,00	3	1
228	2	2	26	1	5	6	3	1	1	186.625,00	1	0,00	0,00	1	5	2	2	1	2	2	1	131.400,00	0,00	2	1
229	2	2	9	1	1	8	2	2	1	64.101,00	1	0,00	0,00	2	5	2	1	1	2	1	1	0,00	0,00	3	1
230	2	1	108	1	5	5	2	1	1	390.050,00	2	0,00	0,00	1	5	2	2	1	2	2	1	78.000,00	0,00	1	1
231	2	2	2	3	3	2	3	2	1	18.800,00	1	0,00	6.400,00	2	5	2	3	3	2	1	2	0,00	0,00	3	1
232	2	2	5	1	1	0	2	2	1	108.469,00	2	0,00	0,00	2	1	2	3	1	2	2	1	0,00	0,00	3	1
233	2	2	82	1	5	12	2	1	1	79.794,00	2	0,00	0,00	2	5	2	2	1	2	1	2	0,00	0,00	2	1
234	2	2	52	3	3	16	1	2	1	344.615,00	2	463.226,00	0,00	2	5	2	2	1	2	2	1	41.670,00	0,00	1	1
235	2	2	2	3	5	0	1	2	2	863.351,00	1	0,00	0,00	1	1	1	3	3	2	2	2	16.500,00	0,00	3	1
236	2	2	21	1	3	1	2	1	1	257.268,00	2	0,00	0,00	2	5	2	1	1	2	2	2	24.840,00	0,00	2	1
237	2	2	0	3	1	0	3	2	1	13.763,00	2	0,00	0,00	1	5	2	3	3	2	2	1	0,00	0,00	2	1
238	2	2	0	3	4	0	3	2	1	79.987,00	1	0,00	71.000,00	1	5	1	3	3	1	2	2	20.000,00	0,00	2	1
239	2	2	0	3	1	0	3	2	1	22.600,00	2	0,00	0,00	1	5	2	3	3	2	2	1	0,00	0,00	4	1
240	2	2	14	1	2	1	2	1	1	17.703,00	2	0,00	0,00	1	5	2	1	1	2	1	1	0,00	593.750,00	2	1
241	2	2	54	3	3	6	1	2	1	53.854,00	2	192.060,00	0,00	1	5	1	2	1	2	2	1	0,00	0,00	1	1
242	2	2	26	1	5	6	2	1	1	49.993,00	2	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	2	1
243	2	2	33	1	3	1	3	2	1	72.066,00	2	0,00	0,00	1	5	2	2	1	2	2	1	45.000,00	0,00	2	1
244	2	2	109	2	5	4	3	2	1	90.019,00	2	0,00	4.500,00	1	4	2	2	1	2	1	2	0,00	0,00	2	1
245	2	2	119	3	5	0	2	1	1	41.591,00	2	0,00	0,00	2	5	1	3	1	2	2	1	0,00	25.300,00	1	1
246	2	2	43	3	1	2	2	1	1	55.920,00	2	0,00	0,00	1	5	2	2	1	2	2	2	0,00	0,00	3	1
247	2	2	45	2	5	3	3	2	1	232.000,00	1	33.600,00	0,00	1	5	2	2	1	2	2	1	17.310,00	28.000,00	1	1
248	2	2	11	3	2	0	2	1	1	48.345,00	2	0,00	0,00	2	5	1	3	1	2	2	2	0,00	0,00	3	1
249	2	2	93	2	5	5	3	1	1	818.726,00	1	0,00	0,00	1	5	2	2	1	2	2	1	598.800,00	0,00	1	1
250	2	2	0	3	1	1	3	2	1	28.080,00	2	0,00	7.700,00	1	1	2	3	3	2	2	2	14.600,00	14.100,00	3	1
251	2	2	0	3	3	2	3	2	1	21.800,00	2	0,00	39.801,00	1	5	1	3	3	2	2	2	0,00	8.946,00	3	1
252	2	2	0	3	1	0	3	2	1	21.000,00	2	0,00	0,00	1	5	1	3	3	2	2	2	0,00	0,00	3	1
253	2	2	42	2	3	6	1	2	2	173.591,00	2	0,00	0,00	2	5	2	2	1	2	2	2	0,00	20.300,00	2	1
254	2	2	0	3	1	0	3	2	1	48.000,00	2	0,00	0,00	1	5	1	3	3	2	2	2	0,00	16.700,00	3	1

255	2	2	13	3	2	0	2	1	2	223.914,00	1	0,00	0,00	2	5	2	3	1	2	1	2	0,00	70.000,00	3	1
256	2	2	12	1	2	2	3	1	1	96.859,00	2	0,00	0,00	2	1	2	3	1	2	2	2	0,00	0,00	3	1
257	2	2	18	1	3	2	2	1	1	466.962,00	1	0,00	6.600,00	2	5	1	1	1	2	2	2	0,00	0,00	3	1
258	2	2	11	1	3	10	3	2	1	128.956,00	1	0,00	2.000,00	1	5	2	1	1	2	2	1	0,00	21.500,00	2	1
259	2	2	0	1	3	2	3	2	1	43.754,00	2	0,00	4.550,00	2	5	2	3	3	2	2	1	26.451,00	4.500,00	3	1
260	2	2	11	3	3	18	3	1	1	89.307,00	1	0,00	0,00	2	5	2	1	1	2	2	1	42.150,00	0,00	3	1
261	2	2	1	3	1	2	3	2	2	11.155,00	2	0,00	0,00	1	5	2	3	1	2	2	1	0,00	0,00	3	1
262	2	2	0	2	5	0	2	2	1	13.467,00	1	0,00	12.450,00	1	5	2	3	1	2	2	2	0,00	0,00	2	1
263	2	2	30	1	3	2	3	2	1	1.154.442,00	2	0,00	9.000,00	2	5	1	2	1	2	2	2	0,00	0,00	3	1
264	2	2	14	3	5	2	2	1	1	134.635,00	1	0,00	59.340,00	2	5	2	3	1	2	2	1	0,00	0,00	1	1
265	2	2	34	1	5	4	1	1	1	90.259,00	2	45.000,00	0,00	1	1	2	2	1	2	2	2	0,00	0,00	1	1
266	2	2	33	1	3	2	2	1	1	82.471,00	2	0,00	0,00	2	5	2	2	1	2	2	1	0,00	0,00	2	1
267	2	2	0	1	1	2	2	1	1	31.485,00	1	0,00	0,00	1	5	2	3	3	2	2	2	37.000,00	0,00	3	2
268	2	2	0	1	5	2	2	1	1	149.143,00	1	0,00	0,00	2	5	1	3	3	2	2	2	40.500,00	0,00	2	2
269	2	2	33	1	4	3	2	2	1	96.760,00	2	0,00	0,00	2	5	2	3	1	2	2	1	0,00	0,00	2	2
270	2	2	12	1	3	2	1	2	1	44.223,00	2	0,00	0,00	0	5	2	1	1	2	1	2	9.000,00	0,00	3	2
271	2	2	0	3	1	0	3	2	1	31.048,00	2	0,00	0,00	1	5	1	3	3	2	2	2	6.000,00	20.200,00	3	2
272	2	2	2	3	2	1	3	2	1	40.077,00	2	0,00	7.000,00	2	5	1	3	1	2	2	1	0,00	0,00	3	2
273	2	2	24	1	3	3	2	1	1	106.644,00	1	0,00	0,00	2	5	2	1	1	2	2	2	0,00	0,00	3	2
274	2	2	1	3	5	0	3	1	1	53.578,00	1	0,00	13.700,00	1	5	1	3	3	2	2	2	18.000,00	0,00	2	2
275	2	2	0	1	1	1	3	2	1	35.750,00	2	0,00	0,00	1	1	2	3	3	2	2	2	0,00	0,00	3	2
276	2	2	4	1	1	12	2	2	1	673.552,00	2	0,00	0,00	2	5	2	3	1	2	2	2	21.000,00	10.500,00	2	2
277	2	2	0	1	5	0	2	1	1	199.329,00	1	0,00	0,00	1	5	1	3	3	2	2	2	4.870,00	25.800,00	2	2
278	2	2	1	3	1	0	3	1	1	164.811,00	2	0,00	3.000,00	1	1	2	3	3	2	2	2	0,00	0,00	3	2
279	2	2	1	1	1	1	1	1	1	75.000,00	1	0,00	0,00	1	1	2	3	3	2	2	2	0,00	10.500,00	3	2
280	2	2	16	2	2	16	2	2	1	784.574,00	2	0,00	0,00	2	5	2	1	1	2	2	2	252.605,00	0,00	2	2
281	2	2	0	2	1	5	2	2	2	180.000,00	1	0,00	0,00	2	1	2	3	3	2	2	1	0,00	49.800,00	2	2
282	2	2	0	1	4	2	3	1	1	180.000,00	1	0,00	0,00	2	5	1	3	3	2	2	1	6.000,00	28.000,00	1	2
283	2	2	0	1	1	1	2	1	1	53.508,00	2	0,00	0,00	1	5	2	3	3	2	2	2	0,00	0,00	3	2
284	2	2	13	3	3	7	2	1	1	71.966,00	2	0,00	0,00	2	5	1	1	1	2	2	2	0,00	0,00	3	2
285	1	1	1	1	1	2	2	1	1	120.000,00	2	0,00	0,00	2	5	2	3	3	1	1	2	0,00	0,00	2	2
286	2	2	0	2	2	0	2	2	2	124.300,00	1	0,00	0,00	2	5	2	3	3	2	2	1	0,00	3.000,00	3	2
287	2	2	9	2	1	2	3	2	1	86.254,00	2	0,00	0,00	1	5	2	1	1	2	2	1	0,00	0,00	3	2
288	2	2	23	2	3	8	3	2	1	249.066,00	1	0,00	0,00	2	5	2	3	1	2	2	2	78.000,00	0,00	2	2
289	2	2	8	1	4	1	2	1	1	77.114,00	1	0,00	0,00	2	1	2	1	1	2	2	2	0,00	17.000,00	2	2
290	2	2	0	1	5	1	3	2	1	66.000,00	2	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	3	2
291	2	2	14	1	3	0	2	2	1	45.455,00	2	0,00	0,00	2	5	2	1	1	2	2	2	0,00	0,00	3	2
292	2	2	11	1	2	4	2	1	1	57.063,00	2	0,00	0,00	1	5	2	1	1	2	2	2	0,00	16.500,00	2	2

293	2	2	0	1	1	0	2	1	1	265.350,00	2	0,00	0,00	1	5	1	3	3	2	2	1	13.546,96	6.500,00	2	2
294	2	2	0	1	1	1	3	2	1	134.500,00	2	0,00	0,00	2	5	1	3	3	2	2	1	0,00	0,00	3	2
295	2	2	4	2	1	0	2	1	1	19.175,00	1	0,00	0,00	1	5	2	3	1	2	2	1	0,00	0,00	3	2
296	2	2	2	3	1	4	2	1	2	190.000,00	1	0,00	0,00	1	5	2	3	3	2	2	2	0,00	12.700,00	3	2
297	2	2	101	3	4	85	1	2	2	319.307,00	2	28.800,00	0,00	2	5	2	2	4	2	2	2	0,00	0,00	1	2
298	2	2	0	3	1	3	3	2	1	52.000,00	2	0,00	0,00	1	5	1	3	3	2	2	2	0,00	0,00	3	2
299	2	2	12	1	5	8	3	2	1	1.352.065,00	1	0,00	0,00	1	5	1	1	4	2	2	1	173.271,00	0,00	3	2
300	2	2	32	1	3	0	2	2	1	213.073,00	2	0,00	0,00	1	5	2	2	4	2	1	2	0,00	0,00	2	2
301	2	2	6	3	1	2	2	2	1	30.162,00	2	0,00	0,00	2	5	2	3	1	2	2	2	0,00	0,00	3	2
302	2	2	21	3	3	11	2	2	2	119.053,00	1	0,00	93.000,00	1	5	2	1	1	2	2	2	0,00	22.000,00	2	2
303	2	2	0	2	1	4	2	2	1	35.160,00	2	0,00	2.400,00	1	5	2	3	3	2	2	1	18.800,00	0,00	3	2
304	2	2	5	2	1	6	2	2	1	123.174,00	2	0,00	16.000,00	1	5	2	3	1	2	1	2	0,00	50.800,00	3	2
305	2	1	31	2	3	1	2	2	2	65.133,00	1	0,00	2.200,00	1	5	1	2	1	2	1	2	0,00	0,00	2	2
306	2	2	4	1	1	0	2	1	1	17.766,00	2	0,00	0,00	1	5	2	3	1	2	2	2	0,00	0,00	3	2
307	2	2	1	2	3	12	3	2	1	591.800,00	1	0,00	0,00	2	5	2	3	3	2	2	2	0,00	26.274,00	2	2
308	2	2	0	3	2	4	3	2	1	54.721,00	2	0,00	0,00	2	5	2	3	3	2	2	2	13.650,00	8.500,00	2	2
309	2	2	0	2	2	8	2	2	1	41.232,00	2	0,00	6.000,00	2	5	2	3	3	2	2	2	0,00	0,00	4	2
310	2	2	3	2	1	6	3	2	1	252.336,00	1	0,00	28.000,00	2	5	2	3	1	2	2	1	6.000,00	0,00	3	2
311	2	1	23	2	3	3	3	2	1	140.619,00	2	0,00	0,00	0	5	2	1	4	2	1	1	0,00	0,00	3	2
312	2	2	2	2	1	0	2	2	1	112.800,00	2	0,00	0,00	2	1	1	3	1	2	2	2	0,00	0,00	3	2
313	2	2	17	1	3	7	3	1	1	76.391,00	2	0,00	0,00	2	5	1	3	1	2	2	2	0,00	0,00	2	2
314	2	2	0	2	3	1	2	2	1	98.658,00	2	0,00	0,00	2	5	2	3	3	2	2	1	0,00	15.200,00	3	2
315	2	2	2	1	2	1	2	2	1	36.479,00	1	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	3	2
316	2	1	13	1	5	4	1	1	1	179.528,00	2	68.250,00	0,00	1	1	2	1	4	2	2	2	0,00	0,00	3	2
317	2	2	0	1	3	0	3	2	1	36.697,00	1	0,00	14.200,00	2	5	1	3	3	2	2	2	0,00	0,00	3	2
318	2	2	0	3	1	1	3	2	1	54.000,00	2	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	2	2
319	2	2	6	1	3	9	3	1	1	240.000,00	2	0,00	110.000,00	1	1	2	1	1	2	2	2	30.000,00	4.000,00	2	2
320	2	2	0	1	1	3	2	2	1	272.000,00	2	0,00	0,00	2	5	2	3	3	2	2	2	3.000,00	8.500,00	3	2
321	2	2	10	1	2	0	2	1	1	65.619,00	1	0,00	0,00	2	5	2	3	1	2	2	1	14.125,00	0,00	3	2
322	2	2	1	1	5	4	3	2	1	93.162,00	2	0,00	0,00	2	5	2	3	1	2	2	1	0,00	0,00	3	2
323	2	2	8	2	1	13	2	2	1	259.600,00	2	0,00	19.750,00	1	1	2	3	1	2	2	2	0,00	2.500,00	3	2
324	2	2	0	1	2	4	2	1	2	142.285,00	2	0,00	2.000,00	2	5	2	3	3	2	2	2	0,00	0,00	3	2
325	1	2	13	1	4	4	2	2	1	110.813,00	1	0,00	0,00	2	5	2	1	1	2	2	2	0,00	4.000,00	3	2
326	2	2	25	2	3	3	3	2	1	60.325,00	2	0,00	0,00	0	5	2	2	1	2	2	2	0,00	0,00	2	2
327	2	2	25	1	5	2	2	1	1	81.484,00	1	0,00	0,00	2	5	2	2	1	2	2	1	0,00	10.500,00	1	2
328	2	2	0	3	5	0	1	2	1	31.291,00	1	0,00	0,00	2	5	1	3	3	2	2	2	0,00	0,00	2	2
329	2	2	0	1	1	1	2	1	1	114.000,00	2	0,00	0,00	1	1	1	3	3	2	2	2	0,00	0,00	3	2
330	2	2	0	3	1	0	3	2	1	18.634,00	2	0,00	16.110,00	2	5	1	3	3	2	2	2	0,00	0,00	3	2

331	2	2	3	1	1	1	2	2	1	128.600,00	2	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	3	2
332	2	2	0	2	3	0	3	2	1	114.076,00	1	0,00	0,00	2	5	2	3	3	2	2	2	0,00	0,00	4	2
333	2	2	0	1	1	0	3	1	1	120.000,00	1	0,00	4.100,00	1	1	2	3	3	2	2	2	0,00	0,00	3	2
334	2	2	15	2	2	4	2	2	1	394.115,00	1	0,00	26.700,00	1	5	2	1	1	2	2	2	21.900,00	0,00	2	2
335	2	2	0	1	1	0	1	2	1	202.428,00	2	0,00	7.000,00	2	5	1	3	3	2	2	2	0,00	0,00	2	2
336	2	2	2	3	1	0	2	1	1	93.610,00	2	0,00	12.780,00	2	5	1	3	3	2	2	2	0,00	0,00	3	2
337	2	2	0	1	1	2	2	2	1	85.861,00	2	0,00	5.700,00	1	5	2	3	3	2	2	2	0,00	0,00	3	2
338	2	2	0	3	5	2	2	2	1	171.591,00	2	0,00	4.000,00	2	5	2	3	3	2	2	1	0,00	0,00	3	2
339	2	2	8	1	4	0	2	1	1	23.688,00	2	0,00	0,00	2	5	2	3	1	2	2	2	0,00	20.450,00	2	2

ANEXO III - CONJUNTO DE TESTES DE REDES NEURAIS - EXEMPLO DE UM RESULTADO

j	A	w1	w2	w3	w4	w5	w6	w7	w8	A*W1	A*W2	A*W3	A*W4	A*W5	A*W6	A*W7	A*W8
1	2	-1,8831	-2,8499	0,2984	-1,5627	-2,2491	-1,7454	-1,4680	-1,9790	-3,766	-5,7	0,5968	-3,125	-4,498	-3,491	-2,936	-3,958
2	2	-2,0469	-1,3046	0,3780	0,6631	0,8473	-1,4772	-0,3533	-0,5954	-4,094	-2,609	0,756	1,3262	1,6946	-2,954	-0,707	-1,1908
3	17	5,9171	1,8912	-4,6696	0,5341	1,2463	1,9918	-4,6731	0,0103	100,59	32,15	-79,38	9,0797	21,187	33,861	-79,44	0,1751
4	1	0,3840	0,4259	0,2256	2,9754	-0,5098	0,4988	-0,4567	0,4121	0,384	0,4259	0,2256	2,9754	-0,51	0,4988	-0,457	0,4121
5	5	0,7572	1,3156	-0,5636	0,4704	-1,2618	0,4419	-0,6023	-1,4041	3,786	6,578	-2,818	2,352	-6,309	2,2095	-3,012	-7,0205
6	5	0,4051	0,7126	-0,6111	-1,9293	-0,0860	0,1761	-0,5675	-1,3770	2,0255	3,563	-3,056	-9,647	-0,43	0,8805	-2,838	-6,885
7	2	0,3733	2,2225	-1,2008	-0,0368	2,2178	1,0665	-0,6860	-1,8050	0,7466	4,445	-2,402	-0,074	4,4356	2,133	-1,372	-3,61
8	1	1,9559	1,8565	0,4710	2,0748	-1,6748	1,6037	-2,2068	-2,0580	1,9559	1,8565	0,471	2,0748	-1,675	1,6037	-2,207	-2,058
9	1	1,5225	-2,4476	1,5702	2,3500	0,7911	-0,5204	-1,4971	-5,0824	1,5225	-2,448	1,5702	2,35	0,7911	-0,52	-1,497	-5,0824
10	1	0,8871	1,7669	-1,2548	1,3864	-0,1701	0,6192	-1,1380	2,4993	0,8871	1,7669	-1,255	1,3864	-0,17	0,6192	-1,138	2,4993
11	2	-0,4237	-0,9765	-2,2526	0,9953	-3,4489	0,8796	0,5059	3,3519	-0,847	-1,953	-4,505	1,9906	-6,898	1,7592	1,0118	6,7038
12	0	0,0832	-1,1558	-0,1442	-0,0494	0,0439	0,0239	-0,1602	-1,6379	0	0	0	0	0	0	0	0
13	0	0,0852	4,3619	0,0405	0,3989	-0,8783	0,0199	0,0081	-2,2191	0	0	0	0	0	0	0	0
14	2	1,1127	-0,3314	0,1344	-3,0788	3,7511	1,0459	-0,4824	0,8866	2,2254	-0,663	0,2688	-6,158	7,5022	2,0918	-0,965	1,7732
15	5	-0,1085	-1,4328	-0,2048	-2,0926	0,2092	0,5301	-0,0245	-0,8961	-0,543	-7,164	-1,024	-10,46	1,046	2,6505	-0,123	-4,4805
16	2	-0,4118	-1,0625	0,1114	-1,6364	-4,4989	1,1898	-0,1873	-1,3586	-0,824	-2,125	0,2228	-3,273	-8,998	2,3796	-0,375	-2,7172
17	1	0,4384	3,8527	0,4400	2,9882	-0,2328	-0,5367	0,1793	0,2201	0,4384	3,8527	0,44	2,9882	-0,233	-0,537	0,1793	0,2201
18	2	-0,6748	-1,4139	0,3059	-3,0443	0,9520	0,7978	-0,7274	0,1387	-1,35	-2,828	0,6118	-6,089	1,904	1,5956	-1,455	0,2774
19	2	0,2606	-1,7765	0,2333	-4,7576	0,0766	0,9600	1,7878	1,7199	0,5212	-3,553	0,4666	-9,515	0,1532	1,92	3,5756	3,4398
20	1	1,1400	-0,2558	0,4034	0,5466	-2,7858	1,8459	1,6044	1,6222	1,14	-0,256	0,4034	0,5466	-2,786	1,8459	1,6044	1,6222
21	1	-0,6473	0,7397	-0,1166	2,3510	0,7006	0,9891	-1,0235	3,9659	-0,647	0,7397	-0,117	2,351	0,7006	0,9891	-1,024	3,9659
22	0	0,4039	0,6684	-0,3202	-1,6524	2,9776	0,1381	-0,3270	-0,6092	0	0	0	0	0	0	0	0
23	0	0,0864	1,8242	-0,1391	1,6731	-1,8692	0,0718	0,0149	-2,9729	0	0	0	0	0	0	0	0
24	5	-0,4101	-2,5894	-0,0701	2,9532	0,4694	-0,1529	-0,3986	1,5039	-2,051	-12,95	-0,351	14,766	2,347	-0,765	-1,993	7,5195
									-6,6358	**102,1**	**13,133**	**-88,88**	**-4,156**	**9,2553**	**48,77**	**-95,17**	**-8,394**

B	R	AA	Ww	AAWw	BB	AAA	Resultado
1,3823	103,48	1	5,5994	5,5994	1,3020	6,9753	**0,9991**
6,2733	19,406	1	-7,8197	-7,82			
-6,7955	-95,67	2,8E-42	-2,9623	-8E-42			**0,9991**
-10,0608	-14,22	6,7E-07	7,9417	5E-06			
4,2295	13,485	1	4,9168	4,9168			
-5,907	42,863	1	3,906	3,906			
3,5996	-91,57	1,7E-40	-2,5157	-4E-40			
6,3446	-2,049	0,11411	-8,1429	-0,929			
			0,9233	5,6733			

LEGENDA DO ANEXO III

Depois de treinada a Rede, é hora de testá-la. Novos padrões constantes do conjunto de testes serão apresentados aos melhores resultados encontrados na fase de treinamento. Nesse momento apenas a etapa *forward* será executada.

A etapa *forward* pode ser sintetizada, como a seguir.

$$i_{p,i} = \sum_j w_{ij} x_{p,j} + \theta_i$$

$$a_{p,i} = \frac{1}{1 + e^{-i_{p,i}}}$$

FORWARD →

$$i_{p,h} = \sum_i w_{hi} a_{p,i} + \theta_h$$

$$a_{p,h} = \frac{1}{1 + e^{-i_{p,h}}}.$$

j = quantidade de neurônios na camada de entrada (neste trabalho igual a 24);

A = informações referentes a uma empresa (padrões), conforme descrito no Capítulo II, Item 2.2. (x_{pj});

w1 a w8 = pesos finais obtidos após 10.000 iterações em cada conjunto de testes (conexões pesos entre a camada de entrada (j) e a camada intermediária (i)) (w_{ij});

A*W1

a = $w_{ij} \cdot x_{p,j}$;

A*W8

B = bias da unidade (i);

R = $i_{p,i} = \sum_j w_{ij} x_{p,j} + \theta_i$;

AA = $a_{p,i} = \frac{1}{1 + e^{-i_{p,i}}}$;

Ww = pesos finais obtidos após 10.000 iterações em cada conjunto de testes (conexões pesos entre a camada intermediária (i) e a camada de saída (h)) (w_{hi});

AAWw = $w_{hi} \cdot a_{p,i}$;

BB = bias da unidade (h);

AAA = $i_{p,h} = \sum_i w_{hi} a_{p,i} + \theta_h$, h = 1;

Resultado = $a_{p,h} = \frac{1}{1 + e^{-i_{p,h}}}$.

ANEXO IV

ÁRVORE DE DECISÃO OBTIDA COM O USO DO WEKA - TESTE 5

J48 pruned tree

```
tempoconta <= 25
|  sóciosrestrições = sim: S (4.0/1.0)
|  sóciosrestrições = nao
|  |  risco = A: S (16.0/2.0)
|  |  risco = B
|  |  |  sóciosrestri5anos = sim: S (7.0)
|  |  |  sóciosrestri5anos = nao
|  |  |  |  sociedadecônjuges = sim: S (9.0/1.0)
|  |  |  |  sociedadecônjuges = nao
|  |  |  |  |  aplicações = 1
|  |  |  |  |  |  funcionarios <= 0: S (2.0)
|  |  |  |  |  |  funcionarios > 0: N (2.0)
|  |  |  |  |  aplicações = 2: S (1.0)
|  |  |  |  |  aplicações = 3: S (1.0)
|  |  |  |  |  aplicações = 4: S (0.0)
|  |  |  |  |  aplicações = 5
|  |  |  |  |  |  experiênciacredito = 1
|  |  |  |  |  |  |  bairro = centro: S (5.0)
|  |  |  |  |  |  |  bairro = outro: N (2.0)
|  |  |  |  |  experiênciacredito = 2: N (1.0)
|  |  |  |  |  experiênciacredito = 3
|  |  |  |  |  |  móveissócios <= 37000
|  |  |  |  |  |  |  sede = proprio: S (2.0/1.0)
|  |  |  |  |  |  |  sede = alugado
|  |  |  |  |  |  |  |  imóveissócios <= 0: S (5.0/1.0)
|  |  |  |  |  |  |  |  imóveissócios > 0: N (3.0)
|  |  |  |  |  |  |  sede = cedido: N (7.0/1.0)
|  |  |  |  |  |  móveissócios > 37000: S (3.0)
|  |  risco = C
|  |  |  restrições5anos = sim: N (2.0)
|  |  |  restrições5anos = nao
|  |  |  |  clientes = fisica
|  |  |  |  |  atividade = comercio
|  |  |  |  |  |  seguroempresa = sim
|  |  |  |  |  |  |  móveissócios <= 4000
|  |  |  |  |  |  |  |  aplicações = 1
|  |  |  |  |  |  |  |  |  tempoconta <= 0: N (3.0)
|  |  |  |  |  |  |  |  |  tempoconta > 0: S (2.0)
|  |  |  |  |  |  |  |  aplicações = 2: N (0.0)
```

```
| | | | | | | | aplicações = 3: N (0.0)
| | | | | | | | aplicações = 4: N (0.0)
| | | | | | | | aplicações = 5: N (5.0)
| | | | | | | móveissócios > 4000: S (2.0)
| | | | | | seguroempresa = nao
| | | | | | | bairro = centro: S (13.0/2.0)
| | | | | | | bairro = outro
| | | | | | | | sociedadecônjuges = sim: S (8.0/2.0)
| | | | | | | | sociedadecônjuges = nao: N (8.0)
| | | | | atividade = industria
| | | | | | sede = proprio: S (2.0)
| | | | | | sede = alugado: N (6.0)
| | | | | | sede = cedido: S (2.0/1.0)
| | | | | atividade = servico
| | | | | | funcionarios <= 1
| | | | | | | móveis <= 0: S (12.0/1.0)
| | | | | | | móveis > 0: N (4.0/1.0)
| | | | | | funcionarios > 1: N (7.0/3.0)
| | | | clientes = juridica
| | | | | bairro = centro: N (3.0/1.0)
| | | | | bairro = outro: S (9.0/1.0)
| | | | clientes = misto: S (0.0)
| | risco = D: S (12.0/1.0)
| | risco = E: S (2.0)
tempoconta > 25: S (134.0/4.0)

Number of Leaves  :     40
Size of the tree :  65
Time taken to build model: 0.66 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         282               92.1569 %
Incorrectly Classified Instances        24                7.8431 %
Kappa statistic                          0.7486
Mean absolute error                      0.124
Root mean squared error                  0.249
Relative absolute error                 36.9415 %
Root relative squared error             60.8764 %
Total Number of Instances              306

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   Class
                 0.975    0.277    0.929     0.975     0.951      S
                 0.723    0.025    0.887     0.723     0.797      N

=== Confusion Matrix ===

   a   b   <-- classified as
 235   6 |   a = S
  18  47 |   b = N
```

ÁRVORES DE DECISÃO - EXEMPLO

Como Redes Neurais é um assunto já bastante explorado na literatura, inclusive com exemplos numéricos de seu funcionamento, como em ADAMOWICZ (2000), será apresentado apenas um exemplo do funcionamento da técnica de Árvores de Decisão.

Seja o conjunto de treino dado pela tabela a seguir.

SEXO	IDADE <26	TEM CARRO	CLIENTE
M	Sim	Não	Verdadeiro
M	Sim	Sim	Verdadeiro
F	Sim	Sim	Falso
M	Não	Sim	Falso
F	Sim	Não	Falso
M	Sim	Não	Verdadeiro
M	Não	Não	Falso
F	Não	Sim	Falso
F	Sim	Não	Falso
F	Não	Sim	Falso

1° PASSO: Cálculo da Entropia relativa aos dados deste conjunto

$$\text{Entropia}(S) = - P_{(+)} \cdot \log_2 P_{(+)} - P_{(-)} \cdot \log_2 P_{(-)}$$

Neste exemplo, temos duas classes: Verdadeiro e Falso.

$$S(Verdadeiro,Falso) = -\frac{3}{10} \times \log_2\frac{3}{10} - \frac{7}{10} \times \log_2\frac{7}{10}$$

$$S(Verdadeiro,Falso) = 0,3 \times 1,737 + 0,7 \times 0,515$$

$$S(Verdadeiro,Falso) = \boxed{0,881}$$

2° PASSO: Cáculo do Ganho de Informação

$$Gain\ (S, A) = Entropia\ (S) - \sum_{v=1}^{N} \frac{|S_v|}{|S|} . Entropia(S_v)$$

- $G(Sexo) = 0,881 - \frac{5}{10} \times S(3,2) - \frac{5}{10} \times S(0,5)$, onde:

$$S(3,2) = -\frac{3}{5} \times \log_2\frac{3}{5} - \frac{2}{5} \times \log_2\frac{2}{5} = 0,971$$

$S(0,5) = 0$, então:

$$G(Sexo) = 0,881 - \frac{5}{10} \times 0,971 = \boxed{0,395}$$

- $G(Idade) = 0,881 - \frac{6}{10} \times S(3,3) - \frac{4}{10} \times S(0,4)$

$$G(Idade) = 0,881 - \frac{6}{10} \times 1 - \frac{4}{10} \times 0$$

$$G(Idade) = \boxed{0,281}$$

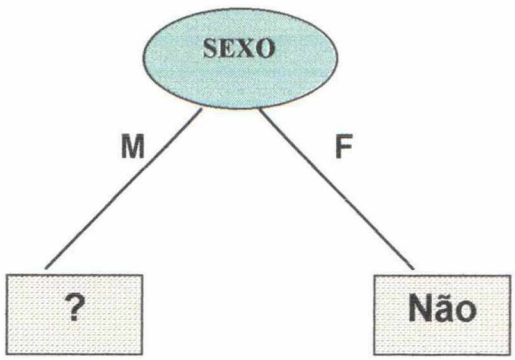
- $G(Carro) = 0{,}881 - \frac{5}{10} \times S(2,3) - \frac{5}{10} \times S(1,4)$

  $G(Carro) = 0{,}881 - \frac{5}{10} \times 0{,}971 - \frac{5}{10} \times 0{,}720$

  $G(Carro) = 0{,}881 - 0{,}485 - 0{,}360$

  $G(Carro) = \boxed{0{,}036}$

Como G(Sexo) > G(Idade)> G(Carro), a Árvore de Decisão é iniciada como a seguir.



3° PASSO: Cálculo do Ganho de Informações considerando os demais atributos

- $G(Idade/Sexo=M) = S(5,5) - \frac{3}{5} \times S(3,0) - \frac{2}{5} \times S(0,2)$

  $G(Idade/Sexo=M) = 1 - \frac{3}{5} \times 0 - 0 = \boxed{1}$

- $G(Carro/Sexo=M) = S(5,5) - \frac{2}{5} \times S(1,1) - \frac{3}{5} \times S(2,1)$

  $G(Carro/Sexo=M) = 1 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0{,}0492$

  $G(Carro/Sexo=M) = \boxed{0{,}570}$

Como G(Idade/Sexo=M) > G(Carro/Sexo=M), a Árvore de Decisão assume o formato a seguir.