ASPECTOS TEÓRICOS DO DATA MINING E APLICAÇÃO DAS REDES NEURAIS EM PREVISÕES DE PREÇOS AGROPECUÁRIOS

LISSANDRA LUVIZÃO LAZZAROTTO; ALCIONE DE PAIVA OLIVEIRA; JOELSIO JOSÉ LAZZAROTTO;

UNIVERSIDADE FEDERAL DE VIÇOSA

VIÇOSA - MG - BRASIL

lislazzarotto@yahoo.com.br

PÔSTER

COMERCIALIZAÇÃO, MERCADOS E PREÇOS AGRÍCOLAS

ASPECTOS TEÓRICOS DO *DATA MINING* E APLICAÇÃO DAS REDES NEURAIS EM PREVISÕES DE PREÇOS AGROPECUÁRIOS

Resumo

A realização de análises relacionadas com determinados bancos de dados pode gerar informações de grande utilidade para os diversos setores da sociedade. Para isso, existem importantes ferramentas de *data mining*, dentre as quais destacam-se as redes neurais artificiais (RNAs). Partindo dessas observações, definiu-se o objetivo geral, que foi avaliar a qualidade de previsões de preços agropecuários, obtidas com o emprego das RNAs. Em termos teóricos e metodológicos, além de uma discussão acerca de relevantes pontos relativos ao *data mining* e às RNAs, foram analisados dois artigos científicos que, usando essas ferramentas, buscaram realizar previsões de preços da soja e da arroba do boi gordo. Dentre os principais resultados, constatou-se que, apesar de existirem algumas limitações, os valores previstos nos dois trabalhos mostraram-se satisfatórios, evidenciando, assim, que as RNAs podem constituir interessante alternativa para a previsão de preços agropecuários.

Palavras-chave: *data mining*, banco de dados, inteligência artificial, economia aplicada, agronegócio

1. Introdução

Ao se analisar o desenvolvimento da sociedade, especialmente nas últimas décadas, verifica-se que é cada vez maior o número de informações geradas. Muitas dessas informações estão armazenadas em grandes bancos de dados que, se devidamente tratados em termos analíticos, podem gerar resultados de significativa relevância para a proposição, elaboração e/ou avaliação dos mais diversos processos e políticas, que cercam os vários setores sociais (produção, serviços, atividades públicas e outros).

A partir da definição dos objetivos que se almeja atingir a partir da análise de certos dados, é possível escolher, dentre diversas opções, a ferramenta de *data mining* que seja a mais adequada. Quando objetiva-se, por exemplo, analisar questões relativas a previsões de

comportamento futuros de certas variáveis, uma das ferramentas que pode ser utilizada, de maneira a assegurar a obtenção de resultados satisfatórios, relaciona-se com as RNAs.

Tomando como referência o setor agropecuário, pode-se afirmar que a análise do comportamento dos preços é uma das condições fundamentais para avaliar, sobretudo, os riscos e a viabilidade econômica das várias atividades que o compõem. Para essa análise, que, em geral, envolve o estudo de certos bancos de dados a partir da utilização de determinadas técnicas (principalmente, com parâmetro matemáticos e estatísticos), pode-se levantar uma importante questão: o uso das RNAs poderia propiciar resultados (previsões) que, de fato, representassem a realidade do setor?

Para buscar respostas a esse problema de pesquisa, foi elaborado este trabalho, cujo objetivo geral foi avaliar a qualidade de previsões de preços agropecuários, obtidos com o emprego das RNAs. Em termos de objetivos específicos, buscou-se atingir três: 1) realizar uma abordagem teórica acerca de importantes aspectos relacionados com a descoberta do conhecimento, dando-se destaque especial para o *data mining*; 2) conceitualizar e demonstrar algumas relevantes aplicações das RNAs; e 3), a partir da análise de dois artigos científicos, analisar, sobretudo, os procedimentos e os resultados obtidos relativos a previsões de preços da soja e da arroba do boi gordo.

Para atingir os objetivos, o trabalho, além desta seção introdutória, contempla quatro seções principais. Na seção dois, são apresentados os fundamentos teóricos, em que são destacados aspectos conceituais e aplicações, tanto do *data mining*, de maneira geral, como das RNAs, de maneira mais específica. A seção três contempla os aspectos metodológicos mais relevantes, que foram utilizados para analisar os dois artigos. A análise detalhada desses trabalhos ocorre na seção quatro. Por fim, as considerações finais deste estudo são apresentadas na seção cinco.

2. Fundamentação teórica

Esta seção está dividida em quatro partes principais. Na primeira, é realizada uma abordagem geral acerca da descoberta do conhecimento e do *data mining*. Na segunda e terceira partes discorre-se, respectivamente, sobre as principais técnicas e alguns exemplos de aplicação do *data mining*. Na quarta parte são efetuadas discussões teóricas relativas às RNAs, com enfoque especial do seu uso no ambiente empresarial.

2.1. Considerações gerais sobre a descoberta do conhecimento e o data mining

A descoberta do conhecimento, que constitui área interdisciplinar específica, conhecida como KDD (Knowledge Discovery in Databases), surgiu em resposta às necessidades de novas abordagens e soluções para viabilizar a análise de grandes bancos de dados, pois os mesmos armazenam conhecimentos valiosos e úteis para os mais diversos processos de tomada de decisões (Romão et al., 2005).

De maneira geral, a KDD pode ser definida como um processo não trivial de identificar padrões válidos, inusitados, potencialmente úteis e, finalmente, compreensíveis em dados (Fayyad et al., citados por Wong e Leung, 2002; Romão et al., 2005).

O conhecimento pode ser extraído diretamente de um banco de dados ou a partir de um armazém de dados, denominado de *Data Warehousing* (Elmasri e Navathe, 2002). Para essa extração, são necessárias ferramentas de exploração, conhecidas como *mineração de dados*, que podem incorporar técnicas estatísticas e/ou de inteligência artificial, capazes de fornecer respostas as várias questões ou descobrir novos conhecimentos (Romão et al., 2005). Portanto, a mineração de dados, ou *data mining*, faz parte do processo de descoberta de conhecimento.

O ato de descobrir padrões úteis em dados recebe, em diversas comunidades, diferentes designações: extração de conhecimento, descoberta de informação, colheita de

informação, arqueologia de dados, processamento de padrão de dados e, inclusive, *data mining*.

O termo *data mining* é muito usado por estatísticos, pesquisadores de banco de dados e comunidades de negócio, constituindo uma das ferramentas mais utilizadas para extração de conhecimento ou informações relevantes, a partir de bancos de dados, nos meios comercial quanto científico ("Data mining overview", 2005; Silberschatz, 1999; Elmasri e Navathe, 2002). Esse termo, a partir do tratamento de grandes quantidades de dados armazenados diretamente em repositórios e por meio da utilização de tecnologias baseadas em ferramentas quantitativas de reconhecimento de padrões, refere-se ao processo de descobrimento de correlações significativas, padrões, tendências, associações e anomalias. Portanto, busca-se, de maneira automática, descobrir regras e modelos estatísticos a partir dos dados ("Data mining overview", 2005; Silberschatz, 1999; Elmasri e Navathe, 2002; Grupo Gartner, citado por Larose, 2005).

A idéia por trás do *data mining* pode causar um certo desconforto devido à ampla gama de objetivos em que o mesmo pode ser usado: uma empresa de varejo interessada em oferecer a melhor oferta para seus consumidores regulares; a receita federal pesquisando transações fraudulentas em remessas de moeda estrangeira; a análise de crédito de um banco, decidindo quais clientes devem receber a próxima mala direta de um novo financiamento; a classificação de clientes de uma operadora de telefonia, sugerindo qual plano se adapta melhor a cada um deles; e outros. Estes são apenas alguns exemplos das inúmeras aplicações do *data mining* (Vessoni, 2005).

Como destacado, o conhecimento obtido a partir de um banco de dados pode ser representado pela definição de regras, que podem ser descobertas por meio do uso de dois modelos: 1) o usuário está diretamente envolvido no processo de descoberta do conhecimento; ou 2) o sistema é responsável por descobrir automaticamente o conhecimento a partir do banco de dados, detectando, assim, modelos e correlações. Entretanto, os sistemas de descoberta do conhecimento podem ter elementos de ambos os modelos: o sistema descobrindo algumas regras automaticamente e o usuário guiando o processo de descoberta de regras (Silberschatz, 1999).

Para descobrir conhecimentos que sejam relevantes, é fundamental o estabelecimento de metas bem definidas. Essas metas, que segundo Fayyad et al. (citados por Romão et al., 2005) são definidas em função dos objetivos associados com a utilização do sistema, podem ser de dois tipos básicos: *verificação* ou *descoberta*. Enquanto na meta de *verificação* o sistema está limitado a verificar hipóteses definidas pelo usuário, na meta de *descoberta* o sistema, de forma automática, deve encontrar novos padrões. A meta do tipo *descoberta* pode, ainda, ser subdividida em *previsão* e *descrição.* A *descrição* procura encontrar padrões, interpretáveis pelos usuários, que descrevam os dados de maneira concisa e resumida, apresentando propriedades gerais, interessantes, dos dados. Na *previsão*, que parte-se de diversas variáveis, é construído um conjunto de modelos, a partir do qual são efetuadas inferências sobre os dados disponíveis, bem como previsões acerca de outras variáveis ou do comportamento de novos conjuntos de dados.

De maneira sintética, a partir da leitura de diversos autores, como Elmasri e Navathe, (2002), "Data mining overview", 2005, Passari (2003), Larose (2005) e Romão et al. (2005), é possível destacar que o *data mining* pode realizar pelo menos uma das seguintes tarefas principais: previsão, sumarização e descrição, classificação, segmentação ou *clustering*, associação e identificação de padrões dentro de séries temporais.

A **previsão**, que lida com comportamentos futuros, envolve a descoberta de um conjunto de informações relevantes para o atributo de interesse. A partir desse conjunto, pode-se prever a distribuição de valores semelhantes ao(s) objeto(s) selecionado(s). Usualmente, a análise de regressão, o modelo linear generalizado, a análise de correlação e as árvores de

decisão têm constituídos as principais ferramentas úteis para a predição de qualidade. Também são usados algoritmos genéticos e redes neurais com bastante sucesso.

Com a **sumarização e descrição** busca-se aumentar o grau de compreensão sobre um fenômeno complexo, representado por grande quantidade de dados de difícil compreensão. Para tanto, são utilizadas, basicamente, técnicas estatísticas descritivas e ferramentas de visualização gráfica. Assim, os resultados do modelo com *data mining* deveriam descrever padrões claros, que são factíveis à interpretação e à explicação intuitiva. Alguns modelos com *data mining* são mais adaptados que outro para a interpretação transparente. Por exemplo, árvores de decisão provêem explicação intuitiva e humanamente amigável dos seus resultados. Por outro lado, redes neurais são, comparativamente, opacas a não especialistas, devido a não linearidade e complexidade do modelo.

A **classificação** constitui a tarefa mais comum de utilização de *data mining*. A partir do exame das características de um grande conjunto de objetos, essa tarefa consiste em colocar cada objeto dentro de uma série de classe ou categoria pré-definidas. Com isso, pode-se tanto entender melhor cada classe no banco de dados, como facilitar a classificação de futuros dados.

A **segmentação ou *clustering*** refere-se à atividade de separar, em grupos homogêneos, uma população heterogênea. Os grupos de registros semelhantes são chamados de *clusters*. A diferença básica entre a segmentação e a classificação é que, enquanto nesta as classes são pré-definidas, na segmentação elas são dinamicamente criadas a partir de similaridades entre os elementos. Portanto, um *cluster* é, em termos gerais, uma coleção de registros que são similares.

A tarefa de **associação** procura descobrir regras para quantificar o relacionamento entre dois ou mais atributos. É usada para determinar afinidades ou ligações entre objetos, consistindo, basicamente, na geração de probabilidades conjuntas (exemplo: quem compra o produto *A* tem um determinado percentual de chances de, também, comprar o produto *B*).

A **identificação de padrões em séries temporais**[1], basicamente, representa uma tarefa cujo objetivo fundamental é identificar comportamentos semelhantes dos dados dentro de posições dessas séries.

Apesar da mineração de dados ser a etapa principal na descoberta do conhecimento, é importante destacar que o processo completo da KDD abrange mais do que a mineração. Ela consiste em seis etapas principais: 1) seleção e limpeza dos dados; 2) pré-processamento; 3) transformação ou codificação de dados; 4) *data mining* e análise; 5) assimilação e interpretação; e 6) avaliação e divulgação das informações descobertas (Elmasri e Navathe, 2002; Wong e Leung, 2002). Além disso, o cumprimento dessas etapas assegura a obtenção do conhecimento útil, derivado dos dados, já que a aplicação de métodos de *data mining* pode ser uma atividade perigosa, que pode conduzir a descoberta de padrões sem sentido ("Data mining overview", 2005).

Em termos operacionais, para efetivação do referido processo, é necessário cumprir três grandes estágios: 1) *pré-processamento*, que consiste, sobretudo, em selecionar os dados mais importantes para o estudo e efetuar as transformações necessárias de modo a serem retiradas as inconsistências e incompletudes dos dados; 2) *data mining,* que, utilizando os dados já preparados no estágio anterior, corresponde à aplicação de métodos (algoritmos) para extrair padrões presentes nos dados; e 3) *pós-processamento*, que está relacionado com a avaliação dos resultados obtidos no estágio de aplicação do *data mining*, visando determinar se algum conhecimento adicional foi descoberto, bem como definir a importância dos fatos gerados. Além disso, para avaliar a qualidade do processo de descoberta de conhecimento, deve-se incluir algumas abordagens, como: exatidão dos resultados (alguma medida da taxa

[1] Séries temporais são séries que possuem regularidade nas observações ao longo de um período de tempo (exemplos: observações diárias, mensais ou anuais).

de acerto), eficiência (tempo de processamento), facilidade de compreensão do conhecimento extraído e outras ("Data mining overview", 2005; Romão et al., 2005).

2.2. Principais técnicas de *data mining*

As mais diversas técnicas empregadas para realizar descobertas de conhecimento, que foram desenvolvidas, sobretudo, pela comunidade de inteligência artificial, tentam encontrar, de maneira automática, regras e modelos estatísticos, que permitem, entre outras coisas, avaliar o comportamento dos dados. De maneira geral, o campo de investigação associado com *data mining* combina idéias de descoberta de conhecimento com a implementação eficiente de técnicas, que possibilitam usá-las em banco de dados muito grandes. (Silberschatz, 1999).

Em termos concretos, as técnicas de mineração de dados estão relacionadas com o uso de algoritmos, que modelam relações ou padrões não-aleatórios (estatisticamente significativos) em grandes bases de dados (Berry e Linoff, citados por Passari, 2003). Nessa mesma linha de pensamento, Romão et al. (2005) ressaltam que as técnicas de *data mining* utilizam dados históricos para aprendizagem, cujo objetivo é realizar uma determinada tarefa particular. Como essa tarefa tem como meta responder alguma pergunta específica de interesse do usuário, é necessário informar qual problema se deseja resolver.

É importante ressaltar que não existe um método de mineração de dados universal, portanto a escolha de um algoritmo particular para uma aplicação é, de certa forma, uma arte (Fayyad et al., citados por Romão et al., 2005). Nessa mesma perspectiva, a partir da leitura dos trabalhos de Gargano e Raggad (1999) e Berry e Linoff (citados por Passari, 2003), podem ser destacados alguns critérios utilizados para a avaliação e a escolha da técnica mais adequada para atingir um determinado objetivo: robustez, grau de automação, velocidade, poder explanatório, acurácia, quantidade de pré-processamento necessário, escabilidade, facilidade de integração, habilidade para lidar com muitos atributos, facilidade de compreensão do modelo, facilidade de treinamento, facilidade de aplicação, capacidade de generalização, utilidade e disponibilidade. Ainda de acordo com os referidos autores, os principais fatores que determinam a escolha da técnica a ser utilizada estão relacionados com preponderância de variáveis categóricas ou numéricas, números de campos, número de variáveis dependentes, orientação no tempo e presença de dados textuais.

Conclui-se, portanto, que não há critérios universais aplicáveis para a escolha e utilização de técnicas de mineração de dados. Isso porque cada técnica possui critérios específicos que devem ser levados em consideração. assim, de acordo com Passari (2003), é também extremamente difícil comparar as técnicas entre si, já que operam de maneira distinta. O autor conclui que a única forma de avaliá-las é por meio da medição de sua habilidade em desempenhar as tarefas para a quais foram construídas.

Apesar de cada técnica de mineração de dados ter sua própria abordagem, elas compartilham algumas características em comum: conforme "aprendem" a partir dos dados de treinamento coletados, ela melhora, gradativamente, a sua performance; e existe sempre uma fase de treinamento, onde o modelo "aprende" os padrões e os relacionamentos (essa fase de treinamento é seguida pela fase implementação, quando o modelo é posto à prova) (Passari, 2003).

Para encontrar respostas, ou extrair conhecimentos interessantes, existem diversas técnicas de mineração de dados. A partir da leitura de alguns trabalhos (Bispo, 1998; Elmasri e Navathe, 2002; Passari, 2003; Wong e Leung, 2002; Romão et al., 2002) que enfocam esse tema, pode-se citar seis técnicas principais relacionadas com *data mining*: indução de regras, redes neurais, algoritmos genéticos, árvores de regressão, lógica nebulosa e *clustering*.

Técnicas de **indução de regras** consistem no uso de ferramental matemático e estatístico, que visam o desenvolvimento de relacionamentos a partir dos dados apresentados.

Tipicamente, são criadas correspondências do tipo "se-então", baseadas em relações causais detectadas nas variáveis. Cada relacionamento "se-então" extraído é chamado de "regra".

As **redes neurais** são técnicas derivadas de pesquisas, na área da inteligência artificial, que utilizam a regressão generalizada. Essas técnicas fornecem *"métodos de aprendizagem"*, pois são conduzidas a partir de amostragens de testes, utilizadas para inferências e aprendizagem iniciais. Com esses métodos de aprendizagem, respostas a novas entradas podem ser passíveis de serem interpoladas a partir das amostras conhecidas. Essa interpolação, no entanto, depende do modelo mundial desenvolvido através do método de aprendizagem.

Os **algoritmos genéticos** estão relacionados com técnicas de otimização, onde se utilizam combinações de processos (exemplo: combinação genética, mutação e seleção natural). Essas técnicas estão associadas, sobretudo, com o conceito de seleção natural.

As **árvores de regressão** são técnicas simples, baseadas na autonomia de uma árvore. Nesse sentido, cada galho particiona, de forma estratégica e sucessiva, os dados em classes e subclasses. A cada divisão, é escolhida a melhor forma de separar e classificar os dados, de acordo com a característica que mais os distingue. Para isso, são utilizadas, também, medidas estatísticas. As árvores de regressão possuem algoritmos não-supervisionados, ou seja, são capazes de processar automaticamente os dados.

Técnicas de **lógica nebulosa** são utilizadas para capturar informações vagas, que, em geral, são descritas na sua forma natural, e convertê-las em um formato numérico, para facilitar as suas análises. Em termos operacionais, essas técnicas utilizam a teoria dos conjuntos nebulosos, que tem mostrado ser muito apropriada para se trabalhar com vários tipos de dados e informações, superando, muitas vezes, os resultados obtidos com o emprego das tradicionais técnicas estatísticas e probabilísticas.

O **clustering** está relacionado com técnicas de *data mining* direcionadas aos objetivos de identificação e classificação. Essa técnica tenta identificar um conjunto finito de categorias, ou *clusters*, para os quais cada objeto de dado pode ser mapeado. As categorias podem ser disjuntas ou sobrepostas e, às vezes, ser organizadas em árvores.

2.3. Algumas aplicações práticas de técnicas de *data mining*

Para mostrar a relevância do uso das técnicas de *data mining* nos mais diversos setores da sociedade, neste item são apresentados oito exemplos de sucesso onde foram aplicadas as referidas técnicas:

a) *a Wal-Mart* constitui uma das maiores cadeias varejistas dos Estados Unidos. É conhecida tanto por sua política de baixos níveis de estoque e ressuprimento constante de produtos (baixos lotes e alta freqüência), como por sua política agressiva com os concorrentes regionais. Utilizando ferramentas de *data mining*, que auxiliam na previsão de cada item transacionado nas lojas da empresa, essa empresa modificou seus sistemas de ressuprimento automático de produtos. Além disso, identificou padrões de consumo, em cada loja, para a escolha do *mix* de produtos a ser ofertado (Rodrigues, 2005);
b) a *ShopKo*, rede varejista americana, utilizou ferramentas de *data mining* para determinar quais produtos eram vendidos por meio da venda indireta de outros produtos. Como resultado, resistiu à concorrência da *Wal-Mart* em 90% dos mercados e, ainda, aumentou suas vendas (Rodrigues, 2005);
c) *o Banco Itaú* costumava enviar mais de um milhão de malas diretas aos correntistas, com uma taxa de resposta de apenas 2%. Com um banco de dados contendo as movimentações financeiras de seus três milhões de clientes, durante 18 meses, e utilizando ferramentas de *data mining*, conseguiu reduzir em um quinto a conta com despesas postais e, ainda, aumentou sua taxa de resposta para 30% (Rodrigues, 2005);

d) *o Bank of America* usou técnicas de *data mining* para selecionar, entre seus 36 milhões de clientes, aqueles com menor risco de dar calote num empréstimo. A partir dos resultados obtidos, enviou cartas oferecendo linhas de crédito para os correntistas cujos filhos tivessem entre 18 e 21 anos e, portanto, precisassem de empréstimos financeiros para ajudar os filhos a comprar o próprio carro, uma casa ou arcar com os gastos da faculdade. Como resultado final, em três anos o banco lucrou 30 milhões de dólares ("Data mining overview", 2005);
e) a empresa *American Express*, a partir da definição de estratégias de *marketing* com o auxílio de técnicas de KDD, fez aumentar as vendas, com utilização de cartão de crédito, em cerca de 20% (Fayyad et al., citados por Romão, 2005);
f) empresas de telecomunicações dos Estados Unidos, a partir da utilização de malas diretas personalizadas com *data mining*, obtiveram reduções da ordem de 45% nas taxas de serviço com novos consumidores (Rodrigues, 2005). Relacionado a esse mesmo setor, segundo o "Data mining overview" (2005), atualmente, existe uma explosão nos crimes contra a telefonia celular, dentre os quais, a clonagem. Assim, técnicas de *data mining* poderiam ser utilizadas para detectar hábitos dos usuários de celulares. Quando um telefonema fosse feito e considerado pelo sistema como uma exceção, o programa poderia fazer uma chamada para confirmar se foi ou não uma tentativa de fraude;
g) no *vestibular PUC-RJ*, utilizando as técnicas de *data mining,* um programa de obtenção de conhecimento, depois de examinar milhares de alunos, forneceu a seguinte regra: *se o candidato é do sexo feminino, trabalha e teve aprovação com boas notas, então não efetiva matrícula.* Uma reflexão justifica essa regra: de acordo com os costumes do Rio de Janeiro, uma mulher com idade para realizar o vestibular, se trabalha é porque precisa, e nesse caso deve ter feito, também, inscrição para ingressar na universidade pública gratuita. Se teve boas notas, provavelmente, foi aprovada na universidade pública, onde efetivará a matrícula. Claro que há exceções: pessoas que moram em frente à PUC, pessoas mais velhas, pessoas de alto poder aquisitivo e pessoas que voltaram a estudar por outras razões. Mas a grande maioria obedece à regra anunciada ("Data mining overview", 2005); e
h) algumas aplicações desenvolvidas pelo *Data Mining* Center (Universidade do Alabama) estão voltadas à utilização de técnicas de mineração de dados para efetuar *previsão de fenômenos naturais.* Dentre os projetos, está o desenvolvimento do AMSU (Advanced Microwave Sounding Unit), que é um radiômetro de microondas utilizado para detectar temperaturas em diferentes níveis da atmosfera. Com base nesse tipo de informação, é possível estimar velocidades de ventos radiais, que, combinadas com outros fatores, podem ser utilizadas para detectar ciclones tropicais (Silva, 2003).

2.4. O uso de redes neurais no ambiente empresarial

Como o objetivo principal deste trabalho está relacionado com a análise da utilização de modelos baseados em redes neurais, visando a previsão de preços, neste item são feitas algumas considerações mais específicas acerca dessas técnicas de *data mining*.

2.4.1. Fundamentos principais das redes neurais

RNAs são sistemas de processamento de informações, compostos por muitos elementos computacionais simples, que interagem por meio de conexões, que recebem pesos distintos. Inspiradas na arquitetura do cérebro humano, as RNAs exibem algumas características, como a habilidade de aprender padrões complexos de informação e generalizar a informação aprendida (Baets e Venugopal, citados por Passari, 2003).

De maneira geral, as RNAs são modelos que relacionam dados de entrada com suas respectivas saídas (Azoff, citado por Ribeiro et al., 2005). A partir da observação de exemplos e seu constante treinamento, obtêm-se uma matriz de pesos, os quais representam as ligações

entre os neurônios de entrada e saída, imitando o que ocorre nas interconexões entre as células nervosas do cérebro humano (Figuras 1 e 2). A grande vantagem do uso das RNAs relaciona-se com a adaptabilidade, que permite seu refinamento e minimização dos erros de previsão (Ribeiro et al., 2005).

As RNAs são compostas de *nós* ou *unidades* (Figura 1), usualmente não-lineares, conectadas por *vínculos* orientados. Um vínculo de unidade *j* para a unidade *i* serve para propagar a *ativação* a_j, desde *j* até *i*. Cada vínculo, também, tem um peso numérico W_{ji} associado a ele, o qual determina a intensidade e o sinal da conexão. Os nós operam em passos discretos, de forma análoga a uma função de dois estágios. No primeiro estágio cada unidade *i* calcula uma soma ponderada de suas entradas (1). O segundo estágio consiste da aplicação de uma função de saída *g* para derivar a saída. Essa função é denominada de *função de ativação* (2) (Russel e Norvig, 2004):

$$in_i = \sum_{j=0}^{n} W_{j,i} a_j \qquad (1)$$

$$a_i = g(in_i) = g\left(\sum_{j=0}^{n} W_{j,i} a_j \right) \qquad (2)$$

onde: *i* e *j* são unidades, a_j é a ativação de saída da unidade *j* e $W_{j,i}$ é o peso no vínculo da unidade *j* até essa unidade.

Portanto, cada nó recebe um ou mais valores de entrada, que são combinados em um único valor a partir do uso de diferentes pesos para cada entrada. Assim, por meio de uma função de ativação, são transformados em um valor de saída. Uma das funções de ativação mais utilizadas é a função logística (3) (Petron, citado por Passari, 2005).

$$f(g) = y = \frac{1}{1+e^{-x}} \qquad (3)$$

onde: *e* é a base do logaritmo natural; *x* é o peso a ser transformado; *y* é o resultado gerado.

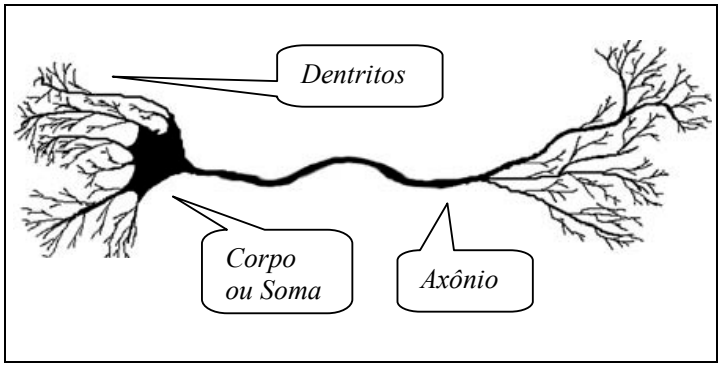

FIGURA 1- A estrutura de um neurônio real. Fonte: adaptado de Larose (2005).

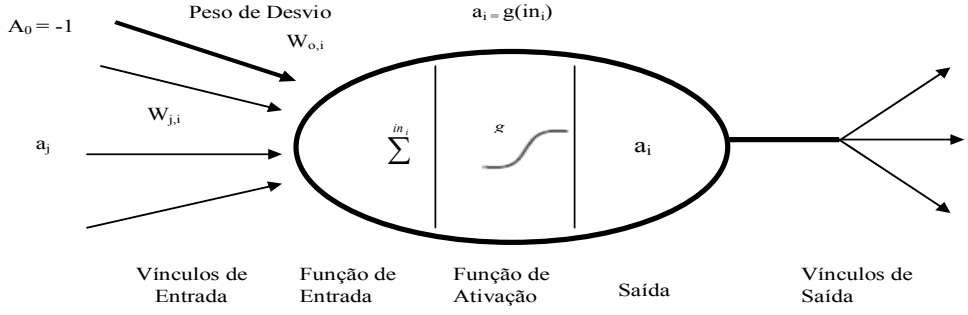

FIGURA 2- Um modelo matemático simples para um neurônio artificial (estrutura de uma RNA). Fonte: (Russel e Norvig, 2004).

Esse funcionamento aparentemente simples de cada neurônio artificial (Figura 2) resulta, após o processamento coletivos de todos os nós, em uma capacidade de execução, com eficiência, de diversas tarefas (Passari, 2003).

As redes neurais podem ser caracterizadas por três propriedades principais: *topologia*, *propriedade computacional* e *propriedade de treinamento* (Baets e Venugopal, Kuo e Xue Almeida, citados por Passari, 2003). A *topologia*, que corresponde à estrutura da rede, refere-se ao número de camadas e nós utilizados. Uma rede neural artificial deve ser composta por pelo menos duas camadas: uma contendo as entradas e a outra contendo as saídas. Em geral são utilizadas, também, uma ou mais camadas intermediárias, caracterizando, assim, as chamadas redes neurais multicamadas: *camada de entrada*, que representa as variáveis de entrada do modelo (essas variáveis devem ser sempre conhecidas); *camada de saída*, que contém um ou mais nós representando os resultados finais do processamento (para uma dada entrada, a rede fornece uma saída correspondente); e *camada intermediárias* ou *ocultas*, que tornam o modelo mais refinado e não-linear, com a capacidade de aprender padrões mais complexos.

A *propriedade computacional* refere-se ao modo pelo qual os nós são ativados e processados, ou seja, corresponde ao "como" e ao "o que" a rede processa.

A *propriedade de treinamento*, que relaciona-se com os aspectos de como a rede aprende, constitui o processo no qual uma série de valores de entrada é apresentada, forma seqüencial, e os pesos da rede são ajustados até que ela reflita a saída desejada. As estratégias de treinamento são divididas em treinamento supervisionado e não supervisionado. O treinamento supervisionado requer a presença de valores de entrada e seus respectivos valores de saída (alvo), a partir dos quais é calculado o erro, utilizado para corrigir o peso das conexões da rede. No treinamento não supervisionado, são apresentados à rede apenas vetores de entrada, não requerendo valores de saída. Nesse caso, a rede agrupa os valores de maior semelhança (*cluster*).

O processo de refinamento dos resultados, mediante o estudo dos erros e envio destes para o início do ciclo, é conhecido como *feedforward backpropagation*. O *feedforward backpropagation* é um algoritmo tradicionalmente usado, que utiliza técnicas de otimização mediante gradiente descendente para o ajuste dos pesos entre conexões. Dessa forma, segue em direção a um ponto de mínimo na curva da superfície de erros (Ribeiro et al., 2005).

Com a característica de poder realizar previsões, além de outras possibilidades, as RNAs representam, portanto, uma importante alternativa aos tradicionais procedimentos estatísticos. Isso porque possui características próprias que, entre outras coisas, facilitam o seu uso em situações onde são exigidas inferências de relações não lineares complexas, entre as variáveis de entrada e de saída, de um modelo previsor (Freiman e Pamplona, 2005).

2.4.2. Algumas aplicações das RNAs

Inúmeras são as possibilidades de aplicações das redes neurais: reconhecimento de padrões (p. ex.: reconhecimento de faces humanas); classificação de dados (exemplo: reconhecimento ótico de caracteres); predição (exemplo.: previsão de séries temporais, como cotações em bolsas de valores); controle de processos e aproximação de funções (exemplo:

robótica); análise e processamento de sinais; filtros contra ruídos eletrônicos; análise de imagens e de voz; avaliação de crédito; entre outras (Freiman e Pamplona, 2005).

No ambiente das organizações, algumas aplicações das redes neurais já são consideradas tradicionais. Nesse sentido, Smith e Gupta (2000) destacam quatro relevantes aplicações:

a) em relação ao ***marketing***, as modernas técnicas consiste em identificar clientes que respondam positivamente a um produto e, assim, direcionar a propaganda para esses clientes. Para tanto, deve-se efetuar uma segmentação do mercado, dividindo-o em grupos distintos de clientes, de acordo com diferentes hábitos de consumo. Essa segmentação pode ser obtida através das RNAs, separando-se a clientela a partir de características básicas: localização geográfica, condição sócio-econômica, poder aquisitivo, atitude em relação ao produto e outras. A partir da segmentação, o *marketing* direto pode ser utilizado para promover as vendas do produto, sendo desnecessária ações intermediárias, como propaganda ou promoção;
b) quanto às **vendas no varejo**, freqüentemente, é necessária ter previsões para que se possa tomar decisões relativas a estoques, contratação de funcionários, preço do produto e outras. Nesse campo, a utilização de RNAs tem tido muito sucesso devido à sua habilidade para considerar, simultaneamente, múltiplas variáveis, como a demanda pelo produto, a capacidade de compra dos consumidores, o tamanho da população e o preço do produto. A previsão de vendas nos supermercados e centrais de atacadistas tem sido muito estudada, e os resultados têm mostrado bom desempenho, quando comparados com aqueles obtidos tanto com o emprego das técnicas de estatísticas tradicionais, como a regressão linear, quanto com a opinião de especialistas;
c) a **área de finanças** tem sido um dos setores com grande aplicação das RNAs visando, sobretudo, efetuar previsões financeiras e de comércio. As RNAs têm sido aplicadas com sucesso em problemas de preço e de *hedge* de derivativos de seguro, previsão de preço futuro para taxa de câmbio e seleção e previsão de desempenho de ações. Atualmente, as RNAs constituem a técnica básica para auxiliar na tomada de decisão com respeito ao risco de crédito, bem como na previsão de falência das corporações. Uma área promissora é a que utiliza RNAs para avaliar as relações entre a estratégia, o desempenho e a saúde financeira das empresas. Além disso, a sua utilização já é significativa para a detecção de fraudes na utilização de cartões de crédito e falsificação de assinaturas em cheques; e
d) na **indústria de seguros** a questão da definição do valor dos prêmios, também, pode ser facilitada pelo uso das RNAs, por meio de previsões da freqüência dos pedidos de indenização. Assim como na área bancária e em outros setores financeiros, aqui se faz necessária a detecção de fraudes a partir da análise de circunstâncias incomuns.

3. Procedimentos metodológicos

Nesta seção são apresentados os principais procedimentos metodológicos utilizados para atingir os objetivos relacionados com o emprego das RNAs em previsões de preços de commodities do agronegócio. Esses procedimentos referem-se ao tipo de pesquisa, ao material de estudo e aos principais aspectos analíticos empregados.

3.1. Tipo de pesquisa e material de análise

Para a realização deste estudo, foram analisados dois artigos técnico-científicos. Esses artigos, a partir da utilização de modelos baseados em redes neurais, buscaram avaliar, sobretudo, a qualidade das previsões de preços a partir do uso dessa técnica de *data mining*. Portanto, o material analítico foi proveniente de pesquisa bibliográfica.

O primeiro artigo refere-se a um trabalho onde busca-se demonstrar possibilidades de aplicação de modelos de RNAs na previsão de preços futuro da soja (Ribeiro et al., 2005). O

segundo artigo trata do uso dessa técnica de *data mining* na previsão de valores pagos pela arroba do boi gordo (Freiman e Pamplona, 2005).

3.2. Principais aspectos analíticos

Sobre os dois artigos selecionados como objeto de estudo deste trabalho, foram analisados cinco aspectos principais:

a) o contexto e a natureza do problema de pesquisa proposto em cada artigo;
b) os principais objetivos almejados pelos autores dos referidos artigos;
c) os procedimentos metodológicos empregados para alcançar os objetivos propostos;
d) os principais resultados alcançados nos dois artigos; e
e) finalmente, as principais conclusões obtidas em ambos artigos.

4. Utilização de RNAs em previsões de preços agropecuários

Esta seção foi elaborada de maneira a conter as principais análises relativas aos dois artigos tomados como objetos analíticos. Para isso, ela foi estruturada em três partes. Na primeira e segunda são efetuadas discussões referentes às previsões, respectivamente, dos preços da soja e da arroba do boi gordo. Na terceira parte são efetuadas algumas análise globais acerca dos dois artigos.

4.1. O caso da previsão do preço futuro da soja

No artigo referente aos preços futuros da soja, os autores partem da idéia inicial de que existem diversos métodos de previsão de preços agrícolas. Dentre esses métodos, incluem-se aqueles baseados em RNAs, pois possuem flexibilidade e habilidade de ajuste para diversas aplicações. Diante disso, o problema de pesquisa elaborado, de certa forma, estava voltado a responder a seguinte questão principal: qual a eficiência em utilizar um modelo, construído sob os pressupostos das redes neurais, para prever os preços futuros da soja?

Como justificativa para a escolha desse problema, destaca-se o fato de o apreçamento de ativos financeiros ser abordado com grande freqüência na literatura. Assim, do ponto de vista cientifico, existe grande potencial de desenvolvimento de modelos matemáticos, o que por si só, justifica o interesse pelo problema. Além disso, hoje existe uma grande diversidade de modelos estatísticos, bastantes sofisticados para determinação de preços. No entanto sua aplicação em mercados agrícolas, muitas vezes, não é possível devido a peculiaridades desses ativos, que estão sujeitos a diversos fatores: condições e regras impostas pelo mercado mundial, protecionismos econômicos, barreiras tarifárias, políticas particulares de cada país e outros.

Partindo desse problema e das justificativas pertinentes, o principal objetivo desse artigo foi construir modelos que antecipassem o comportamento futuro do preço da soja, auxiliando, assim, o processo de tomada de decisões por parte dos agentes econômicos envolvidos com esse produto.

Para atingir esse objetivo, foi implementado um algoritmo no aplicativo Matlab. Em termos de procedimentos metodológicos, primeiramente foram definidas as variáveis mais relevantes, que poderiam afetar o comportamento dos preços pagos pela soja grão: preços de contratos futuros da soja; cotações do dólar; oferta e demanda do produto; número de dias para o vencimento dos contratos futuros; evolução da área plantada; e evoluções das importações brasileiras, exportações mundiais e consumo da soja.

Após a definição das variáveis, definiram-se os procedimentos relacionados com a entrada de dados a serem utilizados na previsão com o emprego das RNAs. Para definir as entradas, foram consideradas as saídas de dois modelos de previsão tradicionais: as médias

móveis simples (MMS) e o alisamento exponencial simples (AES). O MMS consiste em prever o valor para o próximo instante com base nos dados históricos, onde a dimensão da janela de observação do passado é o parâmetro deste modelo. O valor previsto é obtido por meio da expressão (4).

$$M_t = \frac{Z_t + Z_{t-1} + ... + Z_{t-r+1}}{r} \quad (4)$$

onde: r é o intervalo de observações que irá se utilizar e Z_t é o valor da observação realizada no instante t.

O AES realiza uma média ponderada entre os dados históricos e as previsões realizadas anteriormente. Nesse caso, o parâmetro α, que varia entre 0 e 1, determina qual o peso dado para as observações passadas no cálculo da média ponderada. O modelo com parâmetro igual a 1 é o mais simples de todos, pois considera que o valor futuro será exatamente igual à última observação. As previsões são determinadas a partir da expressão (5).

$$Z_t^* = \alpha.Z_z + (1-\alpha).Z_{t-1}^* \quad (5)$$

onde: Z^* é o valor da previsão para o período t; Z é o valor da observação obtida na data t; e α é a constante de alisamento.

Para o MMS foram realizados diversos testes, variando o valor da defasagem (o parâmetro r) da média móvel, afim de determinar o modelo que resultasse num erro quadrático médio (EQM) reduzido e que conseguisse determinar a tendência da série temporal utilizada. Para o AES, a partir da realização de testes, estabeleceu-se uma valor para α igual a 0,3, que correspondia ao parâmetro que suavizava a série relativa aos preços e mantinha erros de previsão reduzidos.

De forma geral, selecionou-se as melhores arquiteturas para o emprego das RNAs por meio de utilização da estatística do EQM (6).

$$EQM = \frac{\sum_{i=j=1}^{n} (x_i - x_j)^2}{n} \quad (6)$$

onde: x_i é o valor observado; x_j é o valor previsto; e n é o número total de observações.

Em relação aos principais resultados, eles foram obtidos, de maneira geral, com cumprimento de quatro etapas:

a. *identificação da função de transferência ou ativação dos neurônios* - o objetivo desta etapa era definir qual das funções de transferência (logarítmica sigmoidal, tangente hiperbólica sigmoidal e linear pura), mediante o emprego do MMS e do AES, resultaria no menor EQM. Para tanto, cada função foi testada com quatro séries: resíduos do MMS, com parâmetro r = 20; série de previsões do AES, com parâmetro α = 0,3; taxas de retorno dos preços da soja; e cotações do dólar. A partir dos resultados, identificou-se que a função logarítmica sigmoidal (logística) constituiu a mais adequada função de transferência para ser utilizada na previsão dos preços em questão;
b. *determinação da primeira variável de entrada nas RNAs* - o objetivo desta etapa foi identificar qual dos modelos (MMS e AES) resultava no menor EQM e, assim, seria a primeira entrada fixa para as RNAs. Para isso, foram realizados testes mediante combinações individuais entre a série obtida com o MMS e as demais séries das variáveis que afetam o comportamento dos preços da soja. Utilizando a série obtida com o AES, realizou-se procedimentos similares. O AES mostrou-se mais adequado, pois os EQMs para esse modelo e para o MMS foram, respectivamente, de 804,74 e 935,27;

c. *determinação da segunda variável de entrada nas RNAs* - como a primeira variável de entrada estabelecida foi a série resultante do emprego do AES, para definir a segunda variável de entrada, foram feitos testes mediante combinações individuais dessa primeira série com as demais séries das variáveis que afetam os preços. Como resultado, identificou-se que a série da variável "prazo para vencimento do contrato futuro da soja" apresentava o menor EQM dentre as variáveis testadas. Esse resultado, de certa maneira, evidencia que o modelo de previsão parece bastante adequado, pois as datas de vencimento dos contratos desse produto estão entre os principais aspectos levados em conta, na previsão de preços, pelos agentes econômicos que atuam no mercado da soja; e
d. *determinação de outras variáveis de entrada nas RNAs* - considerando agora o modelo com duas entradas fixas (AES e os vencimentos dos contratos), foram feitas combinações com outras variáveis que poderiam afetar os preços. Como conclusão, os autores constataram que a demanda e a oferta mundiais deveriam ser consideradas como uma outra importante entrada para rodar o modelo das RNAs. Acrescentando mais variáveis, os resultados do modelo de previsão com RNAs apresentaram desvios maiores em relação aos verdadeiros valores observados.

4.2 O caso da previsão do preço da arroba do boi gordo

Nesse artigo, parte-se da pressuposição de que os modelos estatísticos mais comumente empregados estão relacionados com a metodologia de Box-Jenkins (modelos auto-regressivos) e com o emprego de modelos de regressão, em que podem ser incluídas múltiplas variáveis. Assim, as RNAs podem constituir importante alternativa aos procedimentos estatísticos tradicionais, especialmente nas situações em que são exigidas inferências de relações não lineares complexas entre as variáveis de entrada e de saída de um modelo previsor.

Tomando como base essas observações, os autores do artigo relacionado com os preços da arroba do boi gordo, estabeleceram o seguinte problema de pesquisa: o uso das RNAs pode gerar resultados satisfatórios para prever as cotações mensais da arroba do boi gordo?. Diante disso, o artigo tinha como principal objetivo verificar o poder de previsão das RNAs, em comparação com os tradicionais modelos estatísticos.

Para resolver o problema de pesquisa e atingir o referido objetivo, foram definidas, primeiramente, as variáveis de entrada, bem como o número de conjuntos de observações ("*cases"*) que relacionam a entrada e a saída. Foram definidas sete importantes variáveis de entrada (formam a camada interna), ou seja, que podem influenciar no preço da arroba do boi gordo: mês da previsão; cotação do mês anterior (@-1); cotação de dois meses atrás (@-2); taxa de inflação do mês anterior (IGPM-1); taxa de inflação de dois meses atrás (IGPM-2); e taxa de juros do mês anterior (Selic-1).

Após definida as variáveis de entrada, foi estabelecida a camada intermediária, que inicialmente tinha uma quantidade de neurônios que variava entre 5 e 19, buscando-se, assim, por meio de treinamento e simulação, identificar qual seria a quantidade que traria o melhor resultado. Ainda em relação à camada intermediária, foi estabelecida função de transferência, que deveria ser não linear, de forma a interpolar uma solução para qualquer tipo de problema. A função escolhida foi a sigmoidal (logística) pelo fato de ter uma natureza não linear e, também, devido à sua saída ser formada apenas por valores positivos (essa característica é fundamental, pois os preços do boi gordo são sempre positivos).

Por fim, definiu-se a camada de saída, que era formada por apenas um neurônio tendo em vista que a saída da rede forneceria apenas um valor, que corresponderia a própria previsão do preço da arroba de boi gordo, para um dado mês desejado.

Em termos operacionais, foi também empregado o aplicativo Matlab (interface gráfica). Para fins de padronização dos dados, utilizou-se o *escore z*, que representa o número

de desvios padrão que um valor dista da média do conjunto de valores. Para definir os diferentes meses, foi adotado o sistema binário, ou seja, os diferentes meses assumiam valores 0 ou 1. Visando o treinamento da rede, foram utilizados dados relacionados com as variáveis de entrada (matriz de entrada) e saída (matriz-alvo) correspondentes ao período de abril de 1998 a fevereiro de 2004.

Como teste de validação dos resultados, ou seja, para avaliar o poder de previsão das RNAs, realizaram as previsões para os meses de março a maio de 2004, ou seja, três meses à frente. Esses valores previstos foram, então, comparados com os valores reais ocorridos naquele período, possibilitando, assim, calcular os erros da previsão, que correspondiam a diferença entre os valores previstos e os reais.

Como método estatístico de comparação de resultados obtidos com o emprego das RNAs, foi utilizado o modelo de regressão múltipla (MRM).

Dentre os principais resultados obtidos pelos os autores do artigo analisado, três merecem ser destacados:

a. os melhores resultados para a previsão, mediante o uso das RNAs, foram os obtidos com o treinamento da rede realizado com o algoritmo *backpropagation*, sendo 15 o número ideal de neurônios na camada intermediária;
b. os modelos baseados em RNAs e MRM não são concorrentes, mas complementares. Dependendo da relação entre as variáveis, um deles fornecerá melhor resultado. Quanto maior for a não linearidade entre as variáveis, verifica-se uma tendência natural para as RNAs mostrarem melhor performance na previsão, em relação à regressão; e
c. a partir dos dados apresentados no Quadro 1, pode-se verificar que, dependendo do mês a que se refere a previsão, existem grandes variações nos resultados obtidos com o emprego dos dois modelos.

QUADRO 1 - Valores relativos às cotações reais e previstas (em US$/@) e erros (em US$/@ e em %) obtidos com os usos de RNA e MRM.

Mês/ano	Valores reais (a)	Valores previstos		Erros nas previsões		% erros nas previsões	
		RNA (b)	MRM (c)	RNA (b-a)	MRM (c-a)	RNA (b/a)	MRM (c/a)
mar/04	20,27	19,47	19,84	-0,8016	-0,4300	-3,95	-2,12
Abr/04	20,43	18,00	20,06	-2,4281	2,0581	-11,88	-1,81
Mai/04	19,63	20,20	20,22	0,5747	0,0153	2,93	3,01

Fonte: Freiman e Pamplona (2005).

4.3. Análise global dos casos

Analisando, de forma global, os dois artigos discutidos, pode-se fazer algumas inferências principais:

a. apesar dos dois artigos analisados terem distintos objetos de estudos, ambos utilizaram as RNAs com o mesmo intuito, ou seja, com base em comportamentos históricos das principais variáveis dependentes (preços da soja e da arroba do boi gordo) e explicativas (variáveis determinantes das dependentes), buscaram identificar padrões de comportamento que permitissem efetuar previsões mais adequadas acerca dos dois produtos agropecuários tomados como objetos de estudo;
b. ao avaliar os dois trabalhos, percebe-se, também, que o emprego das ferramentas relacionadas com as RNAs estava voltado a desempenhar tarefas referentes a associações. Nesse sentido, sobretudo, a partir do uso das variáveis explicativas e das funções de transferências, de maneira geral, procurou-se determinar pesos quantitativos que possibilitassem verificar a influência de cada uma dessas variáveis sobre as dependentes;

c. visando atender os objetivos propostos, para fins de treinamento das RNAs utilizadas, ambos os artigos utilizaram séries temporais referentes as principais variáveis de entrada (*inputs*) e saída (*outputs*);
d. quanto à função de ativação, os autores dos dois artigos identificaram que a função logística mostrou-se a mais adequada para ser utilizada nos processos de previsão de preços dos dois produtos. Isso se justifica pelo fato de que esta função é uma das usadas com maior sucesso em diversos processos de previsão. Além disso, ela gera apenas valores positivos, o que é um condição fundamental na previsão de preços, haja visto que não existem preços negativos.

Em termos operacionais, pode-se destacar que nos dois trabalhos analisados os autores, embora tivessem objetivos distintos, para atingir estes, cumpriram as três etapas fundamentais para o desenvolvimento de um processo de KDD. Na etapa de pré-processamento foram definidas as variáveis que realmente poderiam ser significativas para determinar os preços dos produtos. Para isso, realizaram-se testes específicos visando definir tanto o número de camadas de entrada, bem como as variáveis relacionadas a essas camadas. A segunda etapa (*data mining*) consistiu, basicamente, na definição da estrutura neuronal analítica, em que se estabeleceu, por um lado, o número de camadas intermediárias, e de outro a função de ativação do neurônios artificiais, que nos dois casos foi a logística (sigmoidal). Com essa definição cumpriu-se, portanto, a fase de processamento, que permitiu extrair os principais padrões de comportamento dos dados relativos às variáveis explicativas. Por fim, a etapa de pós-processamento esteve relacionado, sobretudo, com a avaliação qualitativa dos resultados (previsões de preços) gerados na etapa anterior. Para tanto, os autores dos dois artigos usaram critérios distintos para efetuar essa avaliação. Enquanto no caso dos preços da soja foi utilizada como medida o EQM, no caso dos preços da arroba do boi gordo utilizaram-se como parâmetros comparativos os resultados obtidos com a aplicação do modelo de regressão múltipla.

5. Considerações finais

O *data mining* é um processo poderoso, que pode depender ou não do usuário, para transformar grandes quantidades de dados em conhecimento. Esse conhecimento poderá auxiliar, sobretudo, nos processos de tomada de decisões nas mais diversas áreas. Apesar dessa grande vantagem em transformar informações em conhecimento, existem também algumas limitações relacionadas, principalmente, com a análise dos resultados gerados. Isso porque, em geral, não podem ser facilmente interpretados por pessoas que não tenham conhecimentos específicos na área, exigindo, assim, profissionais mais especializados.

Dentre as diversas ferramentas do *data mining*, têm-se as RNAs, que são empregadas em muitos setores, visando, entre outras coisas, realizar previsões de comportamentos futuros. Essa ferramenta de previsão, em relação às tradicionalmente usadas, exige do usuário uma maior atenção referente a alguns aspectos: seleção das variáveis de entrada, tipos de função de ativação, definição do número de camadas necessárias para um melhor processamento dos dados e cuidados especiais na interpretação dos resultados gerados, pois as RNAs são, normalmente, de difícil interpretação.

A partir da análise dos dois artigos, é possível inferir que as RNAs podem, de fato, constituir uma interessante alternativa para a previsão de preços da soja e da arroba do boi gordo, haja vista que geraram resultados bastante satisfatórios. Entretanto, uma das grandes limitações na condução de ambos os trabalhos relaciona-se com a falta de um maior embasamento teórico relativo a aspectos microeconômicos, que cercam os referidos preços. Esse embasamento é condição fundamental para melhorar o processo de escolha e definição das variáveis que comporão, principalmente, a camada de entrada, bem como para analisar os resultados gerados.

Os procedimentos metodológicos empregados, bem como os resultados obtidos na condução dos dois artigos, também, possibilitam verificar que, de fato, não existem critérios universais aplicáveis às mesmas ferramentas de *data mining*. Isso porque, embora em ambos os trabalhos o objetivo principal estava voltado à avaliação da qualidade da previsão de preços a partir do uso das RNAs, os autores dos mesmos utilizaram alguns critérios e procedimentos metodológicos bastante distintos, tanto para a obtenção, como para a avaliação dos resultados. Diante dessas evidências, comprova-se que é muito difícil comparar os modelos entre si, uma vez que operam de maneira distinta.

6. Referências Bibliográficas

BISPO, C. A. F. **Uma análise da nova geração de sistemas de apoio à decisão**. São Carlos: UFSCAR, 1998. (Dissertação de mestrado).

DATA mining overview. Disponível em: <http://www.de9.ime.eb.br/~intec/Data%20Mining/Artigos%20de%20Suporte/Overview%20Data%20Mining.pdf>. Acesso em: 10 out. 2005.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**: fundamentos e aplicações. 3.ed. Rio de Janeiro: LTC, 2002.

FREIMAN, J. P.; PAMPLONA, E. de O. Redes neurais artificiais na previsão do valor de commodity do agronegócio. In: Encuentro Internacional de Finanzas, 5. **Anais…**, Santiago, Chile, 2005. 14p.

LAROSE, D. T. **Discovering knowledge in data**: an in introduction to data mining. New Jersey: John Wiley & Sons, 2005.

PASSARI, A. F. L. **Exploração de dados atomizados para previsões de vendas no varejo utilizando redes neurais**. São Paulo: USP, 2003. (Dissertação de Mestrado).

RIBEIRO, C. de O.; SOSNOSKI, A. A. K. B.; WIDONSCK, C. A. Redes Neurais aplicadas à previsão de preços da soja no mercado futuro. In: Congresso Brasileiro de Economia e Sociologia Rural, 43. **Anais...**, Ribeirão Preto: SOBER, 2005. 1 CD-Rom.

RODRIGUES, A. M. Escavando dados no varejo. Disponível em: <http://www.cel.coppead.ufrj.br/fs-busca.htm?fr-varejo.htm>. Acesso em: 08 out. 2005.

ROMÃO, W.; FREITAS, A. A.; PACHECO, R. C. S. Uma revisão de abordagens genético-difusas para descoberta de conhecimento em banco de dados. Disponível em: <http://www.din.uem.br/~wesley/AGFuzzy.pdf>. Acesso em: 13 nov. 2005.

RUSSEL, S. J.; NORVIG, P. **Inteligência artificial**. 2.ed. Rio de Janeiro: Campus, 2004.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de banco de dados**. 3.ed. São Paulo: Makron Books, 1999.

SILVA, M. P. dos S. **Mineração de dados em bancos de imagens**. São José dos Campos: INPE, 2003. (Monografia do Exame de Qualificação do Doutorado em Computação Aplicada).

SMITH, K. A.; GUPTA, J. N. D. Neural networks in business: techniques and applications for the operations researcher. **Computers & Operations Research**, p.1023-1044. Set. 2000.

VESSONI, F. Introdução à Mineração de Dados. (2005). Disponível em: <http://www.mv2.com.br/datamining.doc>. Acesso em: 27 out. 2005.

WONG, M. L.; LEUNG, K. S. **Data mining using grammar based genetic programming and applications**. New York: Kluwer Academic Publisher, 2002.