FACULDADE DE TECNOLOGIA DE SÃO PAULO

LUCIANA MENDES

Data Mining – Estudo de Técnicas e Aplicações na Área Bancária

São Paulo
2011

FACULDADE DE TECNOLOGIA DE SÃO PAULO

LUCIANA MENDES

Data Mining – Estudo de Técnicas e Aplicações na Área Bancária

Monografia submetida como exigência
Parcial para a obtenção do Grau de
Tecnólogo em Processamento de Dados

Orientador: Professor Paulo Roberto Bernice

São Paulo
2011

Dedico esse trabalho ao meu pai que sempre foi a fonte de inspiração da minha vida.

Agradeço a minha família que tanto amo, a minha tia Vera por ter tido toda a paciência, estando ao meu lado em todos os momentos, ao Professor Paulo Roberto Bernice pela orientação e ao meu Pequeno Príncipe e meus amigos que sempre me apoiaram a vencer mais esse desafio.

RESUMO

Descobrir como transformar dados brutos em conhecimento tornou-se um diferencial para as organizações. Bancos de Dados deixaram de ser simples repositórios de dados e passaram a ser explorados a fim de permitir um melhor aproveitamento das informações, ao se extrair e transformar dados brutos em informações valiosas para auxiliar o processo decisório.

Esse trabalho apresenta um estudo sobre as técnicas de Data Mining que é uma parte integral do processo de KDD - *Knowledge Discovery in Databases* e que permite a geração de modelos de dados a partir da aplicação de algoritmos para a extração de padrões de dados. Visando, por fim, apresentar uma forma no qual os métodos do Data Mining possam ser utilizados por instituições bancárias e de crédito a fim de melhorar a qualidade e a eficiência das decisões.

Palavras-chave: data mining, dados brutos, bases de dados, conhecimento, algoritmos, modelos de dados, instituições bancárias.

ABSTRACT

It’s become a differential for organizations to find out how to transform raw data into knowledge. Databases left to be simple repositories of data and started to be investigated order to allow a better use of information, extracting and transforming these data into valuable information helping people to taking decision.

This paper introduces a study on the techniques of Data Mining that is an integral part of the KDD - *Knowledge Discovery in Databases* process so that allows the generation of data models from the application of algorithms to extract data patterns. Aiming, in the end, to prrovide a way in which the methods of Data Mining can be used by banks and credit institutions to improve the quality and efficiency of decisions.

Keywords: data mining, raw data, databases, knowledge, algorithms, data models, banks.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

BB – Banco do Brasil

DM – Data Mining

KDD – *Knowledge Discovery in Databases*

LTDA – Limitada.

MATLAB – Neural Networks Toolbox

WEKA – *Waikato Environment for Knowledge Analysis*

SUMÁRIO

INTRODUÇÃO

O crescimento da coleta e do armazenamento de informações em bancos de dados cada vez maiores e complexos tem gerado a necessidade por parte das empresas de procurar por técnicas que permitam um melhor aproveitamento do investimento realizado. Descobrir como transformar esses dados brutos em conhecimento novo que pode ser usado no processo de tomada de decisão representa um ganho para a empresa, bem como um diferencial competitivo.

Na área bancária, em específico, as instituições financeiras com suas gigantescas bases de dados têm buscado ferramentas capazes de trabalhar com essas informações de forma vantajosa, eficiente e lucrativa. Para isso, tem utilizado métodos inerentes ao processo de Data Mining para se beneficiarem, por meio da extração e transformação desses dados brutos em modelos e padrões que agreguem conhecimento significativo na tomada de decisões.

Todo esse processo será abordado e explicado nas próximas páginas desse trabalho, tendo maior enfoque no que diz respeito ao Data Mining. Iniciará falando de KDD – *Knowledge Discovery in Databases*, ou Descoberta de Conhecimento em Bases de Dados, abordando todas etapas desse processo o qual pertence o Data Mining, em seguida abordará o elemento essencial no estudo do Data Mining: os dados, conceituando seus atributos e tipos de conjuntos de dados. Em seguida, serão explicados os conceitos relacionados às etapas do KDD, para, posteriormente, apresentar algumas das principais técnicas e algoritmos de Data Mining, tais como Árvores de Decisão, Redes Neurais, Descoberta de Regras de Associação, entre outras. Por fim, serão apresentados exemplos de aplicações bem sucedidas do Data Mining em instituições financeiras e de crédito em atividades de análise de crédito, fidelização de clientes, ofertas de produtos, avaliação de riscos e busca por fraudes.

1. KDD - *KNOWLEDGE DISCOVERY IN DATABASES*

1.1. KDD

A todo momento uma grande quantidade de dados é armazenada e, muitas vezes, na tomada de decisão informações essenciais são ignoradas, apesar dos avanços ocorridos nas áreas de Tecnologia de Informação e Banco de Dados. Assim, o crescimento das bases de dados, em tamanho e dimensionalidade, criou a necessidade de aprimorar técnicas e ferramentas capazes de transformar esses dados brutos armazenados em conhecimento.

Essas ferramentas e técnicas para "automatizar e analisar a base de dados de forma inteligente" (LEMOS, 2003) surgem no final da década de 80 em um novo campo de atuação interdisciplinar da informática: o *Knowledge Discovery in Databases – KDD,* ou Descoberta de Conhecimento em Bases de Dados, envolvendo áreas como:

- Banco de dados;
  - Uso de tecnologias específicas a fim de explorar as características dos dados.
- Inteligência artificial;
  - Uso de programas de Inteligência Artificial que oferecem mecanismos para representação do conhecimento, raciocínio e explicações, além de ferramentas para aquisição do conhecimento.
- Estatística e Matemática;
  - Uso da estatística em procedimentos e resultados de testes, bem como de modelos matemáticos para a geração de regras e padrões.
- Reconhecimento de padrões;
  - Uso de modelos cognitivos e paradigmas para a aquisição automática de conhecimento.
- Visualização de dados.
  - Uso de gráficos, figuras e ícones, a fim de permitir a interação entre o processo de descoberta e análise de dados para consulta.

Segundo STEINER (2007), o KDD – *Knowledge Discovery in Databases* é um processo que objetiva analisar, sintetizar e extrair o conhecimento dos dados brutos de grandes bases de dados, por meio de métodos, algoritmos e técnicas dessas áreas, a partir de um conjunto de cinco etapas:

- Seleção dos dados;
- Limpeza de dados;
- Transformação dos dados;
- *Data Mining*;
- Interpretação e Avaliação dos resultados.

As etapas do KDD – *Knowledge Discovery in Databases* têm como principal característica a extração não-trivial de informações implicítas contidas em uma base de dados, já que métodos tradicionais de análise, diante de grande volume de dados relacionados, podem ser ineficazes na detecção de informações úteis, pois tratam apenas de informações explícitas. Essas informações, extraídas pelo processo de KDD, objetivam encontrar relacionamento e padrões entre elementos de dados, ou seja, transformar dados em conhecimento significativo que possam ser relevantes na tomada de decisões.

O processo de KDD inicia-se com a com a definição do domínio da aplicação, ou seja, a coleta de dados que contenham casos e características do problema a ser analisado, e dos objetivos finais da aplicação a serem atingidos. Esses dados são agrupados para, em seguida, passar pelo processo de limpeza de dados e integração de dados. Segundo LEMOS (2003), as dificuldades de integrar bases de dados heterogêneas tornam essa etapa longa, podendo levar até 80% do tempo necessário para todo o processo. Após essas etapas, os dados ainda passam pelo estágio de transformação de dados a fim de que haja o armazenamento adequado para a aplicação dos algoritmos de *Data Mining*, ou Mineração de Dados.

O Data Mining, muitas vezes confundido como sinônimo de KDD, é uma parte integral do processo de extração de conhecimento que permite a geração de modelos de dados a partir da aplicação de algoritmos. Ou seja, um método que enfoca o desenvolvimento de ferramentas especializadas e mais eficientes, capazes de trabalhar com diversos tipos de dados, buscando, assim, padrões de dados a fim de melhorar a qualidade e eficiência das decisões.

Enfim, o processo final é o de Interpretação e Avaliação dos resultados, ou pós-processamento, o qual permite que os dados sejam explorados e validados, bem como os resultados do Data Mining sejam incorporados ao sistema de apoio a decisões.

1.2. Dados

No dicionário online Michaelis, encontramos a seguinte definição para o termo dados:

> Conjunto de material (= informações) disponível para análise. Representação de fatos, conceitos e instruções, por meio de sinais de uma maneira formalizada, possível de ser transmitida ou processada pelo homem ou por máquinas.

Como se pode perceber, os dados representam fatos e conceitos passíveis de serem interpretados por meio do processamento tanto pelo homem como por máquinas, pois permitem a extração de informações e, posteriormente, a análise e transformação dessas em conhecimento.

Assim, tendo em vista que em todos os processos do KDD, os dados são elementos essenciais, logo, antes de prosseguir na explicitação das etapas de seleção, limpeza e transformação dos dados, tratar das questões relacionadas aos dados, tais como conceitos, classificações e tipos, é fundamental, especificamente para o Data Mining, já que o tipo de dado determina quais ferramentas e técnicas podem ser melhor utilizadas.

1.2.1. Conjuntos de Dados

Um conjunto de dados muitas vezes pode ser visto como uma coleção de objetos, de dados, registros, ponteiros, vetores, eventos, casos, observações ou entidades. Por sua vez, objetos de dados são descritos pelo número de atributos que capturam as características básicas de um objeto. Por exemplo, um conjunto de dados é um arquivo, os objetos são registros e o campo corresponde ao atributo.

1.2.1.1. Atributos

Um atributo é uma propriedade ou característica de um objeto que pode variar, seja de um objeto para outro ou de tempo para outro. Atributos também são conhecidos por outros nomes tais como: campo, variável, característica, recurso ou dimensão. Uma forma útil de se especificar o tipo de um atributo é identificar as propriedades de números, ou seja, as

operações utilizadas, que correspondam às propriedades relacionadas do atributo. Quanto à classificação dos atributos, estes podem ser podem ser divididos em categorizados, quando não possuem as caraterísticas dos números, ou numéricos, como demonstrado a seguir:

- Categorizados (Qualitativos)
  - Nominal – Os atributos fornecem apenas informação suficiente para distinguir um objeto de outro. Tem como propriedade a distinção e como operações matemáticas os símbolos = e ≠.
  - Ordinal – Os atributos fornecem informação suficiente para ordenar objetos. Tem como propriedade a ordenação e como operações >, ≥, <, ≤.
- Numéricos (Quantitativos)
  - Intervalar – Os atributos fornecem uma unidade de medida em que diferenças entre os valores são significativas. Tem como propriedade a adição e como operações matemáticas os símbolos + e -.
  - Proporcional – Os atributos fornecem valores em que as diferenças quanto às proporções são significativas. Tem como propriedade a multiplicação e como operações matemáticas os símbolos * e /.

Os atributos também podem são distinguidos pelo número de valores que podem receber. Esses são os seguintes:

- Discretos
  - Um atributo discreto possui um conjunto de valores finito ou infinito, podendo ser categorizados, como códigos postais ou números de ID, ou numéricos, como contadores, desde que esses atributos sejam representados por variáveis de números inteiros.

    Atributo binário é um caso especial de atributos discretos.

- Contínuos
  - Um atributo contínuo possui atributos cujos valores são variáveis de números do tipo real. Geralmente, atributos nominais ou ordinais são dos tipos binários ou discretos, enquanto atributos intervalares ou proporcionais são do tipo contínuo. Entretanto, atributos contadores, que são discretos, também são proporcionais.

- Assimétricos
  - Um atributo é considerado assimétrico quando, apenas a presença – um valor de atributo diferente de zero – é considerado importante. Há casos em que os atributos com valores diferentes de zero são minoria, de modo que focar atributos diferentes de zero é mais significativo e eficiente. Os atributos assimétricos podem ser binários como também dos tipos contínuos ou discretos.

1.2.2. Tipos de Conjuntos de Dados

Os tipos de conjuntos de dados possuem um grande impacto sobre as técnicas de Data Mining no que concerne à escolha dos métodos e algoritmos, podendo ser classificados, segundo TAN (2009), a partir de determinadas características:

- Dimensão ou Dimensionalidade
  - É o número de atributos que os objetos do conjunto de dados possuem. Dados considerados de alta dimensionalidade são conjuntos com centenas ou milhares de atributos. Reduzir a dimensão ou dimensionalidade de um conjunto de dados é algo relevante, pois grandes quantidades de atributos geram dificuldades no processo de Data Mining.

- Dispersão
  - Trabalha com conjuntos de dados com características assimétricas. Assim, dados dispersos, ou seja, apenas os valores diferentes de zero, são armazenados e manipulados, resultando em significativas economias em tempo de computação e armazenamento. Vale ressaltar que alguns algoritmos de Data Mining funcionam apenas com dados dispersos.

- Resolução
  - Há diferentes níveis de resolução para os padrões de dados, pois o tamanho da resolução afeta a visualização dos padrões de dados, podendo encobri-los ou fazê-los desaparecer. Caso seja muito grande, o padrão pode desaparecer, ou se muito pequeno, um padrão pode não ser visível ou ser encoberto por um ruído. Uma imagem com pequena resolução pode perder determinados padrões, enquanto uma

com uma grande resolução pode focar dados específicos não permitir a visualização do todo.

Assim, os conjuntos de dados podem ser divididos em três grupos de tipos de dados, com seus subgrupos, explicitados a seguir:

- Dados em Registros
  - Esse tipo de dados, bastante utilizado pelo processo de Data Mining, trabalha com uma coleção de registros, também chamada de objetos de dados, cada um dos quais consistindo de um conjunto de atributos, ou seja, um conjunto fixo de campos.

    Dados em Registros são normalmente armazenados em arquivos horizontais ou em banco de dados relacionais, que pode ser definido como "uma coleta de dados organizados para servir a muitas aplicações eficientemente pela centralização dos dados e pela minimização de dados redundantes" (HALCSIK, 2007). Evidente que o Banco de Dados é mais do que uma coleção de tabelas, contendo registros, entretanto, para o processo de Data Mining, o banco de dados serve como lugar para localizar registros, uma vez que diversos atributos, ou campos, contidos na base de dados são descartados nos processos de seleção de dados do KDD.

Os Dados em Registros também têm subconjuntos que nada mais são do que tipos diversificados de dados que trabalham com registros, sendo classificados da seguinte forma:

- Dados de Transação
- É um tipo especial de dados em registro. Os registros, também chamados de transação, contém uma coleção de itens cujos campos são atributos assimétricos. Por exemplo, uma compra, vista como uma transação, tem uma coleção de itens comprados, cujos campos têm atributos indicando esse registro de compra e possivelmente, zero, indicando os itens não comprados. Logo, com a maior frequência, os atributos são binários. No entanto, há casos de uso de atributos discretos ou contínuos, por exemplo, indicando o número de itens comprados ou a quantia gasta.

- Matriz de dados
  - É uma variação dos dados em registro que trabalha com atributos numéricos, permitindo que operações de matrizes padrão, possam ser aplicadas na transformação e manipulação os dados. Nesse tipo de dados em registros, trabalha-se com um conjunto de objetos inseridos em uma matriz m por n, em que cada objeto é representado por m linhas, enquanto os atributos são representados por n colunas.

- Matriz de dados dispersos
  - É um caso especial de matriz de dados, em que apenas os valores diferentes de zero são importantes. Nesse caso, os atributos são assimétricos e do mesmo tipo, ou seja, apenas as entradas diferentes de zero de matrizes de dados dispersos são armazenadas.

- Dados Baseados em Grafos

  - Dados com Relacionamento entre Objetos
    - Esse tipo de dados privilegia os relacionamentos entre objetos de dados, que são mapeados por grafos. Segundo TAN (2009), "os objetos de dados são mapeados para nodos do grafo, enquanto os relacionamentos entre objetos são capturados pelas conexões entre os objetos e as propriedades das conexões, como direção e peso." Por exemplo, as páginas da Web conectadas contêm textos, que podem ser considerados objetos de dados. Esses podem ser acessados por meio de conexões existentes entre as páginas, ou seja, por meio do relacionamento existente entre esses objetos.

- Objetos de Dados que são grafos
  - Nesse tipo de conjunto de dados, os próprios objetos de dados são representados como grafos, isto é, os próprios objetos devem ter sub-objetos, denominados estruturas, e que tenham relacionamentos. Um exemplo deste tipo de dados baseados em grafos é a estrutura de compostos químicos. Esta pode ser representada como um grafo, tendo os átomos como nodos e as ligações químicas como conexões.

- Dados Ordenados
  - Trata-se de tipos de dados em que os atributos têm relacionamentos que envolvem ordenação no tempo e no espaço.

Os Dados Ordenados podem ser classificados da seguinte forma:

- Dados Sequenciais
  - É um tipo de dados que funciona como uma "extensão de dados de registros", pois tem um tempo associado a um registro ou a um atributo, permitindo que essa informação torne possível verificar determinado padrão. Por exemplo, em um registro de compras pode haver o tempo em que a transação ocorreu.

- Dados de Sequência
  - Trata-se de um conjunto de dados em que, ao invés de seguirem uma sequência temporal, seguem uma sequência ordenada, ou seja, a posição e a ordem dos objetos são relevantes. Podem ser uma sequência de entidades individuais, como a sequência de palavras ou letras. Um exemplo desse tipo de conjunto de dados pode ser encontrado nas informações genéticas.

- Dados de Séries de Tempos
  - Nesse tipo de dados, cada registro é uma série de tempos, ou seja, uma série de medições feitas no decorrer do tempo. Ao se considerar um conjunto de dados financeiros, por exemplo, objetos que representassem séries de tempos dos preços diários de diversas ações seriam dados de séries de tempo.

- Dados Espaciais
  - São caracterizados por conter, no conjunto de dados, objetos com atributos espaciais, tais como posições ou áreas. Em localizações geográficas, esse tipo de dado pode ser bastante útil ao apresentar dados sobre o clima.

1.3. Etapas do KDD

A seleção, limpeza e transformação dos dados, também conhecidas como pré-processamento, são etapas de preparação essenciais para a aplicação dos algoritmos de Data Mining. A maioria das empresas possui bases de dados, entretanto, nem sempre todos os dados armazenados são necessários para o domínio o qual se pretende aplicar o processo, de modo que a interferência dessas etapas é fundamental.

Assim, a abordagem mais detalhada dessas fases do processo de KDD será apresentada nos tópicos seguintes a fim de mostrar os objetivos de cada uma das etapas, antes de se tratar do processo de Data Mining.

1.3.1. Seleção de Dados

A seleção de dados é a etapa em que é realizada a redução de dados, ou seja, a identificação de quais informações da base de dados devem ser consideradas e utilizadas em todo processo de KDD e, principalmente, no Data Mining. Vale ressaltar que essa fase é relevante, visto que se informações essências forem omitidas, é provável que o modelo ou padrão encontrado tenha precisão limitada.

Logo, os dados aqui selecionados devem conter as informações necessárias a fim de que os resultados finais da análise de dados possam resultar em modelos e padrões eficientes. Segundo NANGIYALIL (2007), o processo de KDD necessita que os dados sejam organizados em uma única tabela de estrutura bidimensional, com os casos e as características do domínio ou problema a ser analisado. Essa organização é necessária para que os dados relevantes para o Data Mining sejam adequadamente identificados e agrupados.

Uma das formas de se realizar essa redução de dados, segundo NANGIYALIL (2007), é pela “Junção de Dados”, que pode ser aplicada pela:

- Junção direta
  - Nesse caso, a partir de uma base de dados transacional, por exemplo, cria-se uma nova tabela com todos os atributos e registros. Os tipos de variáveis e casos inclusos podem ou não ter passado por um processo de análise crítica quanto à contribuição no o processo de KDD.

- Junção orientada

- Ao contrário da junção direta, nesse caso, há uma seleção criteriosa dos atributos que efetivamente possam contribuir no processo de KDD.

  As formas de se aplicar essa seleção podem ser explicadas pelos seguintes métodos:

  - Redução de dados horizontal

    Utiliza métodos como: amostragem aleatória e agregação de informações.

    - Amostragem aleatória

      É um método usado para selecionar um subconjunto dos objetos de dados a serem analisados. Nessa abordagem, um número preestabelecido de registros da base de dados é sorteado, formando uma amostra representativa, de modo que o conjunto final tenha menos registros do que o original, mas com a mesma propriedade do conjunto original de dados.

      Para se determinar o tamanho adequado da amostra utiliza-se "esquemas de amostragem progressiva", isto é, ir de uma amostra pequena de dados até encontrar o tamanho adequado, visto que amostras de menores podem perder padrões ou detectar padrões errôneos, enquanto amostras maiores, apesar de representativas, eliminam a vantagem. Segundo TAN (2009), "o Data Mining usa amostragem porque é custoso demais e consome tempo demais processar todos os dados". Ou seja, reduzir o tamanho dos dados diminuem o custo e o tempo do processo, permitindo a utilização de algoritmos mais eficientes e com bons resultados.

      Vale ressaltar também que há diversos tipos de amostragem aleatórias simples, tais como os casos com ou sem reposição, além dos clusters e da amostragem estratificada. A escolha se dá pelo tipo de dados, de modo que quando o conjunto de dados é de tipos diferentes, a amostragem aleatória simples pode falhar na representação dos tipos de objetos menos frequentes.

    - Agregação

      É um método que elimina atributos, ou seja, há a combinação de dois ou mais objetos em um único, resultando em conjuntos de dados menores. Logo, a agregação torna-se uma vantagem na medida em que essa redução requer menos memória e tempo de processamento, beneficiando o processo de Data Mining na escolha do algoritmo a ser utilizado. Também é vantagem ao fornecer uma visão de alto nível dos dados, o que torna o comportamento de

grupos de objetos ou atributos mais estáveis. A desvantagem consiste na potencial perda de detalhes.

- Redução de dados vertical

  Este método também é conhecido como de redução de dimensão ou dimensionalidade, tendo como objetivo "encontrar um conjunto mínimo de atributos, de tal forma que a informação original seja preservada" (NANGIYALIL, 2007). Ou seja, conjuntos de dados podem ter grande número de características, no entanto, a redução de dimensão ou de dimensionalidade pode levar a um modelo de Data Mining mais compreensível ao eliminar características irrelevantes e reduzir o ruído, permitindo que os dados sejam visualizados mais facilmente. Segundo TAN (2009), "o termo redução de dimensionalidade é muitas vezes reservado para as técnicas que reduzem a dimensionalidade de um conjunto de dados criando novos atributos que sejam uma combinação dos antigos".

  A aplicação desse método pode ser considerada eficiente, pois um conjunto de atributos bem selecionados leva a modelos de conhecimento mais concisos e com maior precisão. Além de que diminui o tempo de processamento e aplicação do algoritmo de Data Mining. Uma das principais vantagens é que a eliminação de um atributo é melhor do que a exclusão dos registros.

  O uso da redução de dados vertical deve levar em conta o seguinte aspecto: a consideração ou não do algoritmo de Data Mining na seleção dos atributos. Caso não se leve em conta, utiliza-se a abordagem Independente do Modelo, e caso se leve em conta a avalição dos resultados obtidos, a abordagem é Dependente do Modelo.

Vale ressaltar que o processo de KDD, tendo em vista a variedade de dados, pode utilizar uma diversidade de métodos de Seleção de Dados tais como eliminação direta de casos, segmentação do bando de dados, redução de valores, entre outros que não são serão abordados neste trabalho, em virtude da extensão do tema.

O essencial é mostrar como esse processo de Seleção pode melhorar o desempenho nas aplicações e algoritmos de Data Mining, processo que será abordado no capítulo seguinte.

1.3.2. Limpeza de Dados

Após a Seleção de Dados necessários para o domínio o qual se pretende aplicar o processo de KDD, tratar da qualidade dos dados é essencial. Segundo TAN (2009), a melhora na qualidade de dados resulta diretamente na melhora da qualidade das análises resultantes, pois elimina a presença de ruídos e dados sem pertinência, perda, inconsistência ou duplicação de dados.

Esse processo de melhora ocorre na etapa de Limpeza de Dados que 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" (LEMOS, 2003).

É sabido que, na maioria das vezes, os dados utilizados advêm de bases de dados que foram criadas para outros propósitos, ou para aplicações específicas. Logo, o processo de Limpeza de Dados permite a detecção e correção dos dados, sendo uma etapa relevante a fim de se evitar problemas de qualidade nas aplicações e algoritmos de Data Mining.

Além da questão da origem das bases de dados, há problemas envolvendo erros humanos, coletas de dados, bem como limitações nos dispositivos de medição que podem afetar os modelos e padrões produzidos. Por exemplo, há dados que se desatualizam rapidamente, como históricos de compra na internet. Entre outros casos, pode haver valores faltando ou até mesmo objetos ilegítimos, duplicados ou inconsistências.

Assim, quanto menos qualidade houver nos dados informados no processo de KDD, maior será a probabilidade de haver erros nos modelos de conhecimento gerados.

1.3.2.1. Métodos de Limpeza de Dados

Dentre os diversos métodos de Limpeza de Dados, este trabalho apresentará uma síntese dos seguintes casos: Correção de valores, Erro de medição e Coleta de dados, Ruídos, Externos, Valores Inconsistentes e Dados duplicados.

- Correção de Valores

  Numa base de dados, é comum faltar um ou mais valores de atributos em alguns objetos. Não é incomum que um objeto não tenha um ou mais valores de atributos. Por exemplo, num cadastro, um CEP ou uma idade podem não estar preenchidos. Entretanto, no processo de Data Mining, esses dados serão

considerados na análise. Assim, o processo de Limpeza de Dados trabalha com essas ausências das seguintes formas:

- Eliminar objetos ou Atributos de dados

  Essa estratégia permite a eliminação dos objetos com dados faltando. No entanto, isso deve ser feito com cuidado, pois o objeto pode conter alguma informação relevante. Outra possibilidade é a eliminação do atributo que tenha muitos valores nulos. No caso de haver muitos objetos nulos ou um atributo com muitos objetos nulos, fazer uma análise mais detalhada é o indicado. Essa estratégia só deve ser utilizada quando campos com valores nulos não forem realmente importantes.

- Eliminar valores faltando

  Em casos em que os objetos ou atributos de dados devem ser considerados, mas há valores nulos, é possível eliminar os valores faltando, ou seja, preencher os dados com valores dos atributos dos pontos mais próximos do ponto com valor faltando. Esse preenchimento pode ser manual ou por meio de algoritmos com valores constantes ou recorrentes.

  Por exemplo, se o atributo for do tipo contínuo, utiliza-se o valor médio do atributo dos vizinhos mais próximos, caso seja um atributo categorizado, o valor de atributo recorrente é usado.

- Erro de medição e coleta de dados

  Consiste em qualquer problema resultante do processo de medição. Caso um valor registrado difira do valor real em alguma extensão, este pode ser considerado um erro de medição. Se ocorrer em atributos contínuos, é denominado erro.

  Já na coleta de dados ou na coleção de dados, trata-se de “erros como a omissão de objetos de dados ou valores de atributos, ou a inclusão inapropriada de um objeto de dados” (NANGIYALIL, 2007).

- Ruídos

  Esse fenômeno, considerado um componente de um erro de medição, ocorre quando há uma distorção de um valor ou adição de objetos ilegítimos. O termo é utilizado em conexão com dados que possuam componente temporal ou espacial.

Como é difícil eliminar ruídos, o Data Mining enfoca o projeto de algoritmos que produzam resultados aceitáveis mesmo com a presença de ruídos nos dados.

- Externos

  São objetos de dados atípicos, ou seja, que têm características diferentes da maioria dos outros objetos de dados no conjunto de dados ou valores de um atributo que sejam incomuns. Mesmo se tratando de dados atípicos, os externos podem representar objetos de dados ou valores legítimos, sendo de interesse do Data Mining. Um exemplo da importância dos externos é na detecção de intromissão na rede e de fraudes.

- Valores Inconsistentes

  Esses dados indicam divergências, ou seja, valores que não coincidem com os dados verdadeiros, devendo ser corrigidos por meio de informações adicionais ou de fontes externas. Por exemplo, quando um CEP não condiz com a cidade indicada.

- Dados duplicados

  Como o próprio nome diz, representa objetos de dados que sejam duplicata, ou quase duplicata, uns dos outros. Trabalhar com a detecção e eliminação de dados duplicados deve levar alguns aspectos em conta, pois se deve tomar cuidado para evitar combinar acidentalmente objetos de dados que sejam semelhantes, mas não duplicados. Por exemplo, duas pessoas distintas com nomes idênticos.

1.3.4. Transformação de Dados

Como visto até agora, o pré-processamento de dados é uma área abrangente que trabalha com diferentes estratégias e técnicas a fim de selecionar objetos de dados e atributos para a análise, bem como para criar ou alterar os atributos de formas complexas. Assim, a transformação de dados que consiste na conversão dos dados para um formato interpretável pelas ferramentas de Data Mining, associada à Integração dos dados, que podem advir de múltiplas fontes, atendem necessidades específicas do Data Mining.

Nesse processo, a Integração de dados é comum ao agregar a objetos existentes mais elementos de outras bases de dados. Caso exista a estrutura de Data Warehouse, é importante verificar a possibilidade de que seja utilizado, já que, segundo LEMOS (2003), o Data Warehouse coleta dados a partir de diversas aplicações de uma organização, integra e organiza os dados em áreas lógicas de assuntos, armazenando as informações de modo que estas possam ser compreensíveis aos tomadores de decisões e para que possam ser aplicadas técnicas de análise e extração de dados. Levando-se em conta que, nesse estágio do processo, os dados estão quase prontos para serem utilizados no Data Mining, com o uso do Data Warehouse as informações contidas podem ser bem aproveitadas.

Como já dito anteriormente, nessa fase os dados necessitam apenas de pequenos ajustes, ou seja, da transformação de dados. Isso ocorre pelo tipo de algoritmo a ser utilizado no processo de Data Mining. Algoritmos de classificação requerem que os dados estejam na forma de atributos categorizados, enquanto de associação requerem que estejam na forma de atributos binários. Assim, a escolha da forma do atributo deve ser aquela que produz o melhor resultado final da análise de dados.

Dentre as transformações, estas podem ser classificadas das seguintes formas:

- Discretização

  Consiste em transformar um atributo contínuo em um categorizado. Nessa tarefa é necessário definir quantas categorias ter e determinar como mapear os valores do atributo contínuo para essas categorias.

- Binarização

  Consiste em transformar atributos contínuos ou discretos em binários. Por exemplo, um atributo que caracterize o sexo de uma pessoa como F, feminino, ou M, masculino, pode ser transformado em binário F = 1 e M = 0.

- Normalização de dados

  Esse processo consiste em ajustar a escala de valores de determinados atributos em pequenos intervalos para que não influenciem o processo de Data Mining. Por exemplo, manter uma escala de -1 a 1 ou de 0 a 1.

2. DATA MINING

2.1. Definições, características e objetivos

> O explosivo crescimento do volume de dados tem gerado uma urgente necessidade de novas técnicas e ferramentas capazes de transformar, de forma inteligente e automática, terabytes de dados em informações significativas e em conhecimento. [...] Em resposta a essa necessidade, surgiu o Data Mining (DM), também chamado de Mineração de Dados.
>
> (SFERRA, 2003, p. 20)

A imensa quantidade de dados disponíveis nos bancos de dados das empresas gerou a necessidade de explorar essa massa de dados a fim de extrair o conhecimento escondido, transformando dados brutos em informações valiosas para auxiliar o processo decisório. Desta forma, o surgimento do Data Mining com seus métodos tornou-se um diferencial.

Desenvolvido nos anos 60, segundo SCAFF (2009), o Data Mining começou utilizando técnicas de análises estatísticas clássicas como correlação, regressão, entre outras. Evoluiu nas décadas seguintes com o surgimento do estudo de lógica *fuzzy e* redes neurais, que foram agregadas ao processo, e finalmente, na década de 90, tomou a forma atual com o uso de técnicas de Inteligência Artificial, utilizadas na solução de problemas empresariais.

Vale ressaltar que o Data Mining, apesar de ter seu surgimento associado a métodos estatísticos, difere de técnicas estatísticas, pois estas buscam padrões hipotéticos, enquanto o Data Mining utiliza os próprios dados para descobrir os padrões. Ou seja, o Data Mining não substitui técnicas estatísticas tradicionais.

Como já dito, além da estatística, o *Data Mining* teve a contribuição de técnicas como redes neurais, método que será abordado posteriormente neste trabalho, e da lógica *Fuzzy*, também denominada Nebulosa, pela sua capacidade de capturar informações vagas, converter em formato numérico e analisar tais dados por meio de conjuntos nebulosos.

O Data Mining, mais do que uma ferramenta de análise de dados ou de análises estatísticas por meio de softwares computacionais, é "uma técnica para determinar padrões de comportamentos, em grandes bases de dados, auxiliando na tomada de decisão (SILVA, 2000)". Como etapa principal do processo de KDD, o Data Mining consiste na aplicação de algoritmos específicos sobre os dados, a fim de abstrair conhecimento. Segundo NANGIYALIL (2007), esses algoritmos utilizam técnicas de aprendizado indutivo sobre

bases de dados e são capazes de extrair conhecimento por meio de exemplos, aplicando métodos interativos por repetidas vezes. Ou seja, esses algoritmos fundamentam-se em técnicas com uma limitação aceitável de eficiência computacional, capazes de produzir uma enumeração particular de padrões.

Quanto aos objetivos, segundo LEMOS (2003):

> "As bases de dados armazenam conhecimento que podem 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 as técnicas convencionais 'falam' à base de dados, enquanto o 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."

Assim, pode-se dizer que seu objetivo principal é a descoberta do conhecimento, por meio de metodologias capazes de extrair informações preditivas e descritivas das bases de dados. Ou seja, extrair informações contidas nos dados, que permitam as empresas mudar as estratégias adotadas, a fim de gerar lucro significativo. Assim, o próprio termo Data Mining, ou Mineração de Dados, advêm da ideia de extrair o "ouro", metal valioso, que no caso da base de dados, é a informação significativa.

2.2. Métodos de Data Mining

O Data Mining, por abranger diversas áreas, tem uma gama de técnicas e ferramentas adequadas ao tipo de problema apresentado. Dentre as técnicas, pode-se classificá-los segundo modelos:

- Preditivos

  São métodos capazes de prever valores futuros ou desconhecidos por meio de algumas variáveis. Isto é, antecipar o comportamento ou valor futuro baseado no conhecimento passado. Dentre os métodos que geralmente utilizam os modelos preditivos, podem-se listar os seguintes:

  - Classificação;
  - Regressões;
  - Detecção de desvios.

- Descritivos

  São métodos capazes de descobrir padrões interpretáveis que descrevam conjuntos de dados. Isto é, encontrar um padrão que consiga explicar os resultados e os valores obtidos. Dentre os métodos que geralmente utilizam os modelos descritivos, podem-se listar os seguintes:

  - Agrupamento ou Clustering;
  - Descoberta de regras de associação;
  - Descoberta de padrões sequenciais.

Além dos modelos, o Data Mining pode ser caracterizado pela escolha e aplicação dos algoritmos a serem aplicados. Dentre os principais métodos existentes, podem-se citar:

- *Regras de Classificação*
- *Regras de Associação*
- Agrupamento ou C*lustering*.

2.2.1. Regras de Classificação

Regras de Classificação, também denominada Classificação ou Classificador, consiste na "tarefa de organizar objetos em diversas categorias pré-definidas" (TAN, 2009). Esta abordagem possibilita a construção de modelos de classificação, ou seja, ser capaz de reconhecer a função que descreva determinada classe a qual um item pertence, a partir de um conjunto de dados de entrada.

Este conjunto de dados de entrada consiste em um conjunto de registros ou objetos, em que cada registro é representado por dois tipos de atributos (x, y):

- ✓ X representa um conjunto de atributos, que podem ser do tipo discreto ou contínuo;
- ✓ Y representa um atributo especial y, do tipo discreto, designado como *rótulo de classe ou classe*.

Assim, a tarefa da classificação é descobrir os padrões que classificam determinados registros numa certa classe de um conjunto definido de classes.

Tendo esses conceitos bem claros, é possível dizer que um modelo de classificação também pode servir como ferramenta explicativa para se distinguir entre objetos e classes diferentes, como para prever o rótulo da classe de registros não conhecidos.

Vale ressaltar que as técnicas de classificação são mais apropriadas para prever ou descrever conjunto de dados com categorias nominais ou binárias. Dentre os principais métodos classificadores, pode-se citar os seguintes, que serão melhor abordados nos próximos tópicos:

- Árvores de Decisão;
- Baseados em Regras;
- Redes Neurais.

Cada um dos métodos classificadores utiliza um algoritmo denominado de Algoritmo de Aprendizagem que objetiva construir modelos com boa capacidade de generalização e/ou que prevejam com precisão os rótulos de classes de registros não conhecidos previamente.

Esses algoritmos trabalham com os seguintes elementos:

- Conjunto de treinamento

  Fornece-se um conjunto de registros com rótulos de classes conhecidos;

- Modelo de classificação

  É construído para relacionar o conjunto de atributos com os rótulos de classes.

- Conjunto de teste

  São registros com rótulos de classes desconhecidos.

Os algoritmos de Aprendizagem funcionam da seguinte forma: um conjunto de treinamento, ou seja, registros com rótulos conhecidos são fornecidos; esse conjunto é usado para construir um modelo de classificação, em seguida, esse modelo é aplicado a um conjunto de teste, que são registros com rótulos de classes desconhecidos.

Após a aplicação do algoritmo, faz-se necessário avaliar o desempenho do modelo de classificação. Essa avaliação é baseada na contagem dos registros de testes previstos pelo modelo, verificando-se quais estão corretos ou não. Assim, o desempenho do algoritmo é calculado pela métrica de precisão e taxa de erro. Vale ressaltar que a maioria dos algoritmos de classificação busca modelos que, quando aplicados ao conjunto de testes, tenham maior precisão ou a menor taxa de erro.

2.2.1.1 Árvores de Decisão

Árvore de Decisão é um dos métodos de classificação, segundo NANGIYALIL (2007), indicado quando a meta do Data Mining é a predição de saídas ou a classificação de dados. Esta técnica é indicada quando se objetiva categorizar dados de arquivos, bem como gerar regras que podem ser facilmente entendidas, explicadas e traduzidas para linguagem natural.

Esse método tem como vantagem considerar os atributos mais relevantes, podendo ser aplicado em grandes conjuntos de dados. Isso ocorre, pois, os atributos são classificados em um número finito de classes, escolhidos e apresentados por ordem de importância, permitindo ao usuário ter uma melhor compreensão do resultado do algoritmo aplicado e uma melhor visão no processo de decisão.

Essa escolha de atributos pela ordem de importância se dá pela forma como a Árvore de decisão é construída. Segundo LEMOS (2003), Árvore de Decisão consiste em:

- Nós ou nodos

  Representam os atributos. Podem ser divididos em:

  - Um nodo raiz ou primeiro nó é o atributo mais importante.
  - Nodos internos são os nodos subsequentes, por relevância, que possuem exatamente uma aresta chegando e duas ou mais saindo.

- Arestas

  Recebem os valores possíveis para estes atributos.

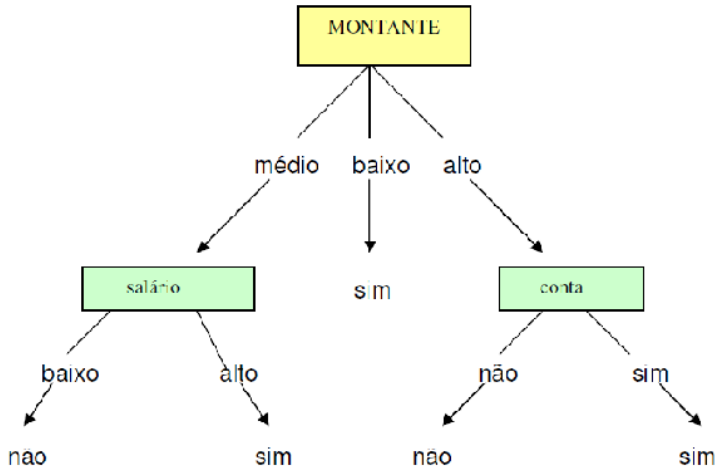
- Folhas ou Nodos folhas

  Correspondem às diferentes classes a que pertencem às entidades.

Assim, uma Árvore de Decisão tem rótulos de classe representados pelas folhas, enquanto o nodo raiz e os outros nodos internos "contêm condições de testes de atributos para separar registros que possuam características diferentes" (TAN, 2009). Esse processo de classificação ocorre, a partir do nodo raiz. Aplica-se uma condição de teste ao registro, seguindo a ramificação pelas arestas baseado no resultado do teste, até que se chegue ao nodo folha. Ou seja, o resultado final da Árvore de decisão está na determinação do rótulo da classe.

Por exemplo, a imagem abaixo mostra um modelo de Árvore de Decisão com poucos atributos, mas que contém todas as características necessárias a esse método:

Figura 1 – Exemplo de Árvore de Decisão



Fonte: LEMOS, (2003).

Como se pode ver, um problema de classificação é resolvido por meio de uma série de questões organizadas sobre os atributos do registro de teste. No exemplo dado, tem-se:

- ✓ o nodo raiz, denominado Montante;
- ✓ os nodos internos, denominados Salários e Conta;
- ✓ as arestas direcionadas com as características: Médio, Baixo, Alto que levam a outros nodos internos;
- ✓ o nodo folha, ou rótulo de classe, com as classificações Sim ou Não.

Assim, por meio de questões sobre as características, cada resposta leva a questão seguinte até que se chegue a uma conclusão sobre o rótulo da classe do registro.

Como já dito, o exemplo dado contém poucos atributos, entretanto, uma Árvore de Decisão pode conter grandes volumes de dados. Assim, após a construção da Árvore de Decisão, esta pode ser podada e otimizada, a fim de reduzir o número de nós internos e a sua complexidade. Podar uma árvore consiste em remover alguma estrutura já construída pelo método.

Essa estratégia é utilizada com da decomposição dos problemas. Segundo LEMOS (2003), a função da Árvore de decisão é "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)." Assim, durante a construção e aplicação da Árvore realiza-se constantemente uma análise e comparação para que se obter novas classes.

Outra vantagem das árvores de decisão é que estas permitem a derivação de regras, que são geradas durante o trajeto da raiz até uma folha de árvore. Derivar regras, também denominado Classificação baseada em Regras, é um método útil quando as árvores são muito grandes.

2.2.1.2. Classificação baseada em Regras

A Classificação baseada em Regras, denominada também de Indução de Regras ou Derivação de Regras, trata do processo de analisar uma série de dados, detectando tendências ou padrões sobre os dados. Essa técnica para classificar registros utiliza um conjunto de regras do tipo "SE... ENTÃO" para gerar regras de previsão. O "SE" funciona como uma regra de condição, especificando alguns valores de atributos preditivos, enquanto a regra "ENTÃO" estabelece a previsão, ou seja, prevê um valor para um determinando atributo.

A construção de um algoritmo de Classificação baseado em Regras identifica relacionamentos entre os atributos de um conjunto de dados e o rótulo da classe, extraindo, assim, um conjunto de regras.

Segundo TAN (2009), para que esse método seja eficiente, as regras descobertas devem satisfazer três propriedades:

- Fazer previsões corretas

  Esse quesito diz respeito à lógica, pois se, na maioria das vezes em que a parte "SE" da regra for verdadeira, a parte "ENTÃO" também deve ser verdadeira;
- Ser compreensível ao usuário

  As regras devem representar conhecimento em alto nível de abstração, ao invés de equações matemáticas complexas;
- Ser útil na tomada de decisões

  A regra deve expressar conhecimento novo.

Quanto aos métodos para extrair regras de classificação, estes podem ser classificados em:

- Métodos diretos

  Extraem regras de classificação direto dos dados. O algoritmo utilizado extrai das regras a classe por vez em conjuntos de dados que contenham mais de duas classes.

- Métodos indiretos

  Extraem métodos de classificação de outros modelos de classificação, tais como árvores de decisão ou redes neurais.

2.2.1.3. Redes Neurais

Redes Neurais são técnicas computacionais que simulam a estrutura e o funcionamento do cérebro humano, por meio de um modelo inspirado na estrutura neural dos organismos inteligentes, isto é, capazes de adquirir conhecimento por meio da experiência. Uma Rede Neural apresenta semelhanças com o cérebro humano por possuir neurônios artificiais.

Tal qual à estrutura do cérebro humano, no qual as redes neurais biológicas possuem suas estruturas interligadas por nós, possuindo uma camada de nós de entrada e de saída, capazes de executar processos através de seus, o método de Redes Neurais é composto de um conjunto interconectado de nodos e ligações direcionadas, ou seja, neurônios artificiais e suas conexões. Vale ressaltar que os nós ou nodos de uma arquitetura de rede neural são também conhecidos como neurônios.

O neurônio artificial possui um ou mais sinais de entrada e de saída. Vale ressaltar que os sinais de saída também podem funcionar como sinal de entrada a outros neurônios, de modo que a conexão de um conjunto de neurônios pode gerar procedimentos complexos.

A função de um neurônio artificial, segundo NANGIYALIL (2003), é:

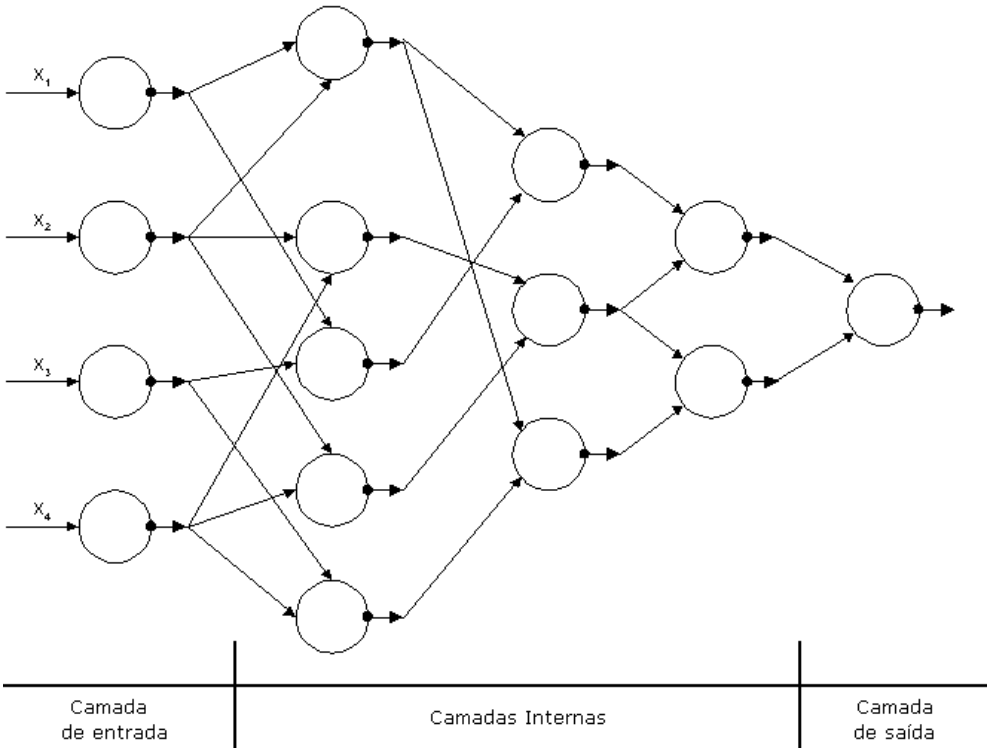
- Avaliar os valores de entrada;
- Calcular o total de valores de entrada combinados;
- Comparar o total com um valor limiar;
- Determinar o que será a saída.

Segundo TAFNER, citado por LEMOS (2003):

> "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)." (2003, p.68)

Tal qual um neurônio que se comunica com outros formando sinapses, as redes neurais recebem e retransmitem informação. Ou seja, um gama de neurônios artificiais recebem informações de outros o tempo todo, retransmitindo-as também. A figura abaixo demonstra como os neurônios de uma Rede Neural funcionam:

Figura 2 – Rede Neural



Fonte: SANDRI (2009).

A complexidade dessa forma de comunicação torna necessária a classificação da rede de neurônios em três camadas, composta por tipos de neurônios, segundo sua função:

- Camada de Entrada

  É a camada inicial da rede de neurônios, onde ficam os neurônios responsáveis pela apresentação de padrões à rede;

- Camadas Intermediárias (Internas) ou Ocultas

  É a camada onde os neurônios responsáveis pela maior parte do processamento, por meio de conexões ponderadas, extraem as características;

- Camada de Saída

  É a camada onde se localizam os neurônios responsáveis pelo resultado final.

Pela sua complexidade, as Redes Neurais têm algumas desvantagens quando aplicadas no processo de KDD, pois geram resultados de difícil compreensão ao usuário. Entretanto, é aplicável de forma eficaz em dados que contêm ruído.

Dentre os modelos de Redes Neurais, os principais abordados neste trabalho são:

- Perceptron

  É o modelo de Rede Neural mais simples. Segundo TAN (2003), o Perceptron consiste de dois tipos de nodos:

  - Nodos de entrada

    Usados para representar os atributos de entrada, transmitem o valor que recebem para a ligação que sai sem executar qualquer transformação.

  - Nodo de saída

    Usado para representar a saída do modelo, é um dispositivo matemático responsável pelo resultado final. Ou seja, responsável pela soma ponderada das entradas, subtração do termo de tendência e finalmente pela saída que depende do sinal da soma resultante.

Para LEMOS (2003), Perceptron é o modelo de rede neurais que aprende conceitos. Por meio de exemplos apresentados aos neurônios da camada de entrada, esse método pode "aprender" a classificar os dados atribuídos como verdadeiro ou falso.

- Multicamadas

  Esse tipo de Rede Neural possui uma estrutura mais complexa do que a do modelo Perceptron. Enquanto o Perceptron é uma rede neural de apenas uma camada, porque possui apenas uma camada de nodos que executa operações matemáticas complexas: a camada de saída. O modelo Multicamadas pode conter diversas camadas intermediárias entre suas camadas de entrada e saída. Essas camadas são chamadas de camadas ocultas, tendo nodos internos: nodos ocultos.

  Outra complexidade consiste no fato de esse método "usar tipos de funções de ativação além da função de sinal. Estas funções de ativação permitem que os nodos ocultos e de saída produzam valores de saída que sejam não lineares nos seus parâmetros de saída" (TAN, 2009).

  Logo, esse tipo de Rede neural é capaz de modelar relacionamentos mais complexos entre as variáveis de entrada e saída. O Algoritmo utilizado nesse caso é o de Retropropagação.

2.2.2. Regras de Associação

Esse é um método bastante útil quando se tem grandes bancos de dados e o proposto for a descoberta de relacionamentos escondidos. Entretanto, as Regras de Associação devem levar em consideração as seguintes questões:

- Pode ser computacionalmente custoso descobrir padrões a partir de um conjunto grande de dados de transações.
- Alguns dos padrões descobertos podem acontecer simplesmente ao acaso, não devendo ser considerados nas Regras de Associação por serem falsos.

A Regra de Associação consiste na seguinte expressão de implicação: $X \rightarrow Y$. Nessa relação, X e Y, antecedente e consequente, respectivamente, são conjuntos disjuntos de itens, $X \cap Y = ₵$. Como já dito, a descoberta de um padrão de implicação pode existir pelo acaso,

logo, é necessário medir a veracidade da Regra de Associação descoberta. Essa verificação, segundo TAN (2009), é realizada por meio dos seguintes elementos:

- Suporte

  Determina a frequência na qual uma regra é aplicável a um determinado conjunto de dados. É uma medida importante, pois é utilizado para eliminar regras sem interesse, bem como verificar a existência de uma regra que tenha baixo suporte por coincidência.

- Confiança

  Determina a frequência na qual os itens em Y aparecem em transações que contenham X. Ou seja, mede a confiabilidade da inferência feita por uma regra. Caso se tenha uma determinada regra $\mathbf{X} \rightarrow \mathbf{Y}$, o nível de confiança determina a probabilidade. Nesse caso, quanto maior a confiança, maior a probabilidade de que Y esteja presente em transações que contenham X. A confiança também fornece uma estimativa da probabilidade condicional de Y dado X.

2.2.2.1. Descoberta de Regras de Associação

Ao tratar de Descoberta de Regras de Associação, a entrada de dados deve consistir de atributos binários denominados itens, visto que, em uma transação, a presença de um item é mais importante do que sua ausência. Logo, um item é um atributo binário assimétrico, já que apenas os dados com a presença de um valor de atributo diferente de zero são considerados padrões frequentes importantes e de interesse.

Assim, dado um conjunto de transações, encontrar todas as Regras de Associação, segundo CARVALHO (2000), requer o uso de alguns algoritmos capazes de extrair associações entre os diversos itens da base de dados.

Além desses aspectos, o resultado dos algoritmos de Regras de Associação deve levar em conta o suporte e a confiança, pois as regras de associação geradas devem ter suporte maior ou igual aos limites do suporte, bem como a confiança maior ou igual aos limites de confiança. Ou seja, devem conter os itens muito frequentes na base de dados, com um grau de relação grande entre eles. Um problema de Regras de Associação busca por todas as regras baseadas nos fatores: suporte e confiança.

Como normalmente a quantidade de dados é grande, uma estratégia adotada pelos algoritmos de Regras de associação é a divisão do problema em subtarefas, que, segundo TAN (2009), são as seguintes:

- Geração de Conjuntos de Itens Frequentes

  Objetiva encontrar todos os conjuntos de itens que satisfaçam o limite do suporte;

- Geração de Regras

  Objetiva extrair todas as regras de alta confiança dos Conjuntos de itens frequentes gerados. Ou seja, gerar regras fortes.

Vale ressaltar que os requisitos computacionais para a Geração de conjunto de itens frequentes são geralmente mais custosos do que a Geração de regras.

Além da divisão de subtarefas, há um algoritmo considerado relevante na Descoberta de Regras de associação:

- Algoritmo Apriori

  Gera regras de associação por meio da abordagem de níveis para, em que cada nível corresponde ao número de itens que pertencem ao consequente (Y) da regra. Ou seja, faz uma varredura sobre o arquivo, a fim de gerar todos os conjuntos de combinações de valores que pertençam ao consequente. Assim, são extraídas todas as regras de confiança alta que tenham apenas um item no consequente, e posteriormente, essas regras são utilizadas para se gerar novas "regras candidatas". Caso o nível da confiança for baixo, as regras contendo o item são descartadas. Entretanto, se tiver um alto nível de confiança, há uma fusão das regras, gerando uma regra candidata. O exemplo abaixo é uma demonstração simples dessas implicações:

  - $\{1, 3, 4\} \rightarrow \{2\}$ e $\{1, 2, 4\} \rightarrow \{3\}$
  - $\{1, 4\} \rightarrow \{2, 3\}$ é a regra candidata.

Segundo CARVALHO (2000), o algoritmo Apriori, caso trabalhe com uma grande base de transações, pode gerar um número imenso de combinações de padrões em determinados limites de suporte e confiança, muitos dos quais poderiam não ser de interesse. Logo, uma das abordagens para definir critérios é por meio de auxílio de argumentos estatísticos.

2.2.2.2. Padrões de Sequências

Padrões de sequências consistem de listas ordenadas de elementos ou eventos, extraídos a partir de um conjunto de entrada de dados que podem ser conjuntos de dados com atributos contínuos, categorizados e até mesmo sequências e grafos.

Esse método caracteriza-se por sua extensão e pelo número de eventos ocorrentes, em que a extensão corresponde ao número de elementos presentes na sequência, enquanto que uma sequência x é uma que contém x elementos.

Segundo TAN (2009), dado um conjunto de dados D contendo uma ou mais sequências de dados, o termo refere-se a uma lista ordenada de eventos associada a um único objeto de dados.

Por exemplo, num Banco de Dados Cliente, há uma sequência que corresponde ao histórico de compras de um determinado cliente, os quais os elementos ou a transação corresponde ao conjunto de itens comprados por esse cliente em um tempo x, sendo que o evento ou item são os produtos. Padrões de sequências podem ser usados para encontrar padrões de páginas visualizadas na Web, sequência de disciplinas feitas por um aluno de graduação, sequências de genoma, entre outros.

2.2.2.3. Padrões de Subgrafos

Este método trabalha com entidades mais complexas, sendo, computacionalmente, mais custoso no processo de Data Mining por causa da escala exponencial do espaço de pesquisa. É responsável pela derivação de um conjunto de subestruturas comuns entre a coleção de grafos. Esse tipo de dados, grafo, consiste de uma estrutura de dados matemática utilizada para representar os relacionamentos entre um conjunto de entidades. Isto é, um grafo é uma estrutura composta por um:

- Conjunto de vértices;
- Conjunto de arestas que conectam pares de vértices.

Segundo TAN (2009), cada aresta é denotada por um par de vértices (*vi*, *vj*), em que *vi*, *vj* pertencem ao conjunto de vértices. O rótulo de classe de cada entidade pode ser representado pelo vértice *vi*, enquanto que as arestas (*vi*, *vj*) associadas ao rótulo descrevem o relacionamento entre as entidades.

Assim, os subgrafos correspondem às estruturas de grafos extraídas dentro de um conjunto de grafos principal. Por exemplo, um grafo com 6 vértices e 11 arestas pode conter um subgrafo com 4 vértices e 4 arestas. Encontrar esses grafos não é uma tarefa fácil, visto que é necessário calcular o suporte e o nível de confiança de tais estruturas a partir de um conjunto de grafos, pois apenas os grafos desconectados, isto é, se tiver apenas arestas não direcionadas, são descartados.

Logo, pela tarefa de minerar subgrafos frequentes ser bem complexa, já que cada entidade é um vértice e pode ter até d – 1 arestas para outros vértices, a extração de subgrafos pode ocorrer por meio do algoritmo do tipo Apriori, de modo que uma coleção de grafos são mapeadas como itens, sendo o número de arestas as transações.

A aplicação dessa técnica de Data Mining pode ser bastante útil para aplicações de Computação em Redes, Bio-Informática, Química Computacional e Mineração na WEB. Em uma aplicação de Química Computacional, por exemplo, os grafos são representados pela estrutura dos conjuntos químicos, tendo os átomos ou íons representados pelas vértices e as ligações entre átomos ou íons, pelas arestas.

2.2.4. Agrupamento ou Clustering

O Agrupamento ou Clustering é um método de Data Mining que segmenta dados em grupos diferentes cujos itens são semelhantes. Segundo NANGIYALIL (2007), esta técnica agrupa informações homogêneas de grupos heterogêneos entre os demais, apontando o item que melhor represente cada grupo. Desta forma, é possível perceber as características de cada grupo, dividindo-os objetos em grupos semelhantes e grupos diferentes.

Segundo TAN (2003), essa análise de grupos permite a divisão de determinado conjunto de dados em grupos, ou clusters, significativos e úteis. Caso o objetivo da técnica de Agrupamento seja grupo com significados, então os clusters devem capturar a estrutura natural dos dados. Entretanto, há casos em que a análise de grupos é utilizada apenas um ponto inicial para outros propósitos, como resumo de dados.

Esse método agrupa objetos baseado apenas em informações encontradas nos dados que descrevem os objetos e seus relacionamentos. O objetivo é que os objetos dentro de um grupo sejam semelhantes entre si, isto é, relacionados, e diferentes de outros objetos de outros grupos. Supondo, por exemplo, que se tenha o objeto clientes com vários atributos. A

diferenciação dos clientes em grupos por atributos pode revelar grupos diferentes, ou seja, clientes agrupados por comportamentos de compras, tendendo a ter comportamentos bem semelhantes como bem diferentes. Quanto maior a homogeneidade dentro de um grupo e maior a diferença entre grupos, melhor ou mais distinto será o agrupamento.

Assim, um agrupamento consiste em um conjunto inteiro de grupos e, segundo NANGIYALIL (2007), pode ser distinguido em diversos tipos de Agrupamento:

- Particional

  É uma divisão do conjunto de objetos de dados em subconjuntos não interseccionados, ou seja, cada objeto de dado deve constar em um subconjunto.

- Hierárquico

  É um conjunto de grupos aninhados organizados como uma árvore. Cada grupo, denominado nodo, na árvore é a união de seus subgrupos, ou filhos, e a raiz da árvore é o grupo contendo todos os objetos.

Vale salientar que o Data Mining, tendo em vista a imensa quantidade de técnicas e algoritmos, pode utilizar uma diversidade de métodos que não são serão abordados neste trabalho, em virtude da extensão do tema.

O objetivo é apenas apresentar alguns conceitos essenciais que podem ser aplicados em diversas áreas, dependendo simplesmente do tipo de problema apresentado a fim de se selecionar o melhor método a ser aplicado.

2.3. Avaliação e Interpretação de Resultados

O processo pós-Data Mining é a Avaliação e a Interpretação de Resultados, quando os resultados do processo podem ser analisados, verificando eventual necessidade de retornar algum estágio anterior, ou a interpretação de padrões pode dar suporte à informação.

3. APRESENTAÇÃO DE ESTUDOS DE TÉCNICAS E APLICAÇÕES DE DATA MINING NA ÁREA BANCÁRIA

3.1. Relacionando as Técnicas de Data Mining e sua Aplicabilidade

Esse capítulo tem o propósito de apresentar alguns estudos e exemplos nas quais determinadas técnicas de Data Mining apresentadas anteriormente foram testadas ou aplicadas, essencialmente, por instituições bancárias e de crédito. Assim, conhecer como e quais métodos de Data Mining, em específico, essas instituições têm utilizado como ferramentas na busca por padrões de consumo e tendências de mercado é um aspecto considerado relevante nesse trabalho.

Logo, os tópicos seguintes abordarão casos e estudos a fim de demonstrar como determinadas instituições bancárias, tais como qualquer empresa, utilizam o Data Mining para aumentar sua rentabilidade e diminuir os riscos de perda financeira, bem como para minimizar os efeitos de possíveis fraudes, por meio da manipulação inteligente dos dados.

3.2. Exemplos de Aplicações do Data Mining na Área Bancária

Instituições bancárias e de crédito com suas gigantescas bases de dados têm utilizado a tecnologia em busca de ferramentas capazes de aumentar a qualidade e a eficiência na tomada de decisões no que concerne às práticas de fidelização dos clientes, ofertas personalizadas de produtos e aplicações financeiras, análise de crédito e de risco e prevenção de fraudes, por meio de técnicas inerentes ao Data Mining, principalmente por métodos como: Descoberta de Regras de Associação, Redes Neurais e Árvores de Decisão.

Isso ocorre, pois, segundo THOMÉ (2009):

> "O que o mercado procura hoje são maneiras ou técnicas que permitam tirar maior proveito do investimento feito na coleta e no armazenamento de montanhas de dados sobre o seu negócio. O desafio está em descobrir e extrair conhecimento novo a partir dos dados, que este conhecimento seja útil e que ao ser usado no processo de tomada da decisão, possa representar um diferencial competitivo e um ganho real para a empresa."

Assim, tornar grandes bases de dados em informações relevantes e que sejam capazes de gerar lucro ou minimizar perdas é um investimento que tem sido frequente nas grandes instituições bancárias e de crédito por apresentarem este diferencial competitivo, já que os

riscos e a concorrência nessa área são considerados alto, por envolverem transações com volumes financeiros efetivamente elevados. Logo, por menor que seja a margem de ganho na área bancária, está é considerada uma oportunidade a ser explorada.

Por exemplo, caso uma instituição financeira tenha interesse em aumentar sua margem de segurança em concessões de empréstimo, a técnica de Árvores de Decisão pode ser aplicada para classificar futuros clientes, bem como descobrir regras a partir dos padrões gerados por esse método.

Assim, partindo de um histórico de empréstimos de clientes que a instituição bancária possui em seu banco de dados, são construídos dois conjuntos de dados: um de treinamento e um de testes, por meio de um algoritmo de classificação de Árvores de Decisão. O algoritmo gerará padrões e aquele que tiver a menor taxa de erro, será o padrão considerado capaz de classificar futuros clientes como: confiáveis, ou seja, possíveis clientes com baixo risco de empréstimo ou como não confiáveis, que são os possíveis inadimplentes.

Dentre os exemplos reais de como o uso do Data Mining pode ser diferencial no setor bancário, destacam-se os casos do *Bank of America* e *AIB Bank*, que, segundo SCAFF (2005), utilizaram-se de ferramentas de *Data Mining* para estabelecer um relacionamento mais direto com clientes, criando linhas de créditos apropriados aos perfis estabelecidos, bem como no estabelecimento de novos negócios no setor financeiro. Essas instituições chegaram a alcançar, em 3 anos, lucros de até 30 milhões de dólares. Outro caso bem sucedido é o do Banco de Montreal que, por meio do Data Mining, pôde extrair o comportamento de seus clientes e tomar decisões estratégicas de negócio.

O Data Mining também pode ser uma ferramenta benéfica nas estratégias de vendas de produtos ao permitir a identificação do perfil financeiro do cliente, gerando assim estratégias para maximizar o lucro. Esse processo envolve a técnica de Descoberta de Regras de Associação capaz de revelar afinidades ou aversões de determinados produtos por clientes.

No setor bancário brasileiro, o Banco Itaú, segundo SCAFF (2005), melhorou suas estratégias com os clientes, ao utilizar ferramentas de *Data Mining*, para gerenciar as movimentações de 3 milhões de clientes. Ao invés de enviar mais de um milhão de malas diretas aos correntistas, com uma taxa de resposta de 2%, ao implementar métodos de Data Mining aumentou a taxa de resposta para 30%, reduzindo um quinto das despesas nessa área.

Como visto, instituições financeiras comumente utilizam aplicações de Data Mining envolvendo métodos como Árvores de Decisão e Descobertas de Regras de Associação já citados nesse trabalho para análise de crédito, avaliação de risco, potencial transações

fraudulentas em cartões de crédito, dentre outras que podem ser bastante eficientes com outras técnicas de Data Mining.

Entretanto, é importante salientar que o uso do Data Mining:

> "não elimina a necessidade de conhecimento e entendimento do negócio e a compreensão dos dados a serem minerados, nem mesmo substitui analistas e pesquisadores da área (ou gestores de negócios) [mas...] proporciona aos usuários meios para encontrar tesouros de informações que permitam detectar tendências e características disfarçadas, confirmar a necessidade de estudos de novas relações" (SFERRO, 2003, p.32).

Logo, mesmo com o uso das melhores técnicas de Data Mining, o papel do analista, gestores e pesquisadores não pode ser desconsiderado, já que a aplicação e a análise adequada desses dependem de bons profissionais. Também é relevante não haver confusão entre Data Mining e técnicas complexas de consulta que permitem bons resultados, mas dependem da formulação prévia de hipóteses, bem como extração manual de informações pelo próprio usuário.

3.3. Estudo realizado por sobre Análise de Crédito Bancário

Este tópico tem por objetivo apresentar o estudo realizado por Eliane P. Lemos, em sua Dissertação de Mestrado denominada: "Análise de Crédito Bancário com o Uso de Data Mining: Redes Neurais e Árvores de Decisão". Nesse trabalho, técnicas de Data Mining específicas, como Redes Neurais e Árvores de Decisão são aplicadas a uma base de dados real no intuito de verificar um padrão de análise de crédito bancário para pessoas jurídicas, ao se verificar o grau de confiabilidade das empresas, classificando-as como: adimplentes ou inadimplentes.

A proposta de gerar classificadores da base de dados capazes de sinalizar quais os procedimentos adequados para ceder crédito, com maior margem de segurança, objetivava auxiliar no processo decisório em novas concessões de crédito.

A base de dados utilizada pela pesquisadora consistia em dados reais de 339 micro e pequenas empresas, clientes do Banco do Brasil, da agência Guarapuava, no Estado do Paraná. Destas, 266 eram adimplentes e 73 inadimplentes. Todos tiveram seus dados cadastrais analisados na agência por estarem dentro do padrão de micro e pequenas empresas,

já que a análise de crédito de empresas de médio e grande porte é de responsabilidade de uma divisão específica do banco.

Os critérios utilizados pela pesquisadora constituíram da análise de um formulário de cadastro contendo diversas variáveis contidas, tais como: Existência de restrições em nome da empresa, Tempo de conta no Banco do Brasil (BB), Faturamento Bruto Anual, Bens Imóveis e Móveis, Histórico da Conta Corrente, Risco atribuído pelo Banco, Crédito junto a Fornecedores, Conceito na Praça, Pontualidade, entre outros.

As técnicas para implementação foram: Árvores de Decisão e Redes Neurais, objetivando, assim, realizar um estudo comparativo dos resultados obtidos em cada um dos métodos, verificando-se qual oferecia o melhor padrão e a menor porcentagem de erros para o contexto do presente trabalho.

3.3.1. Aplicação real do Método de Árvores de Decisão

Nesse estudo, a pesquisadora utilizou o software computacional WEKA, ou Waikato Environment for Knowledge Analysis, para implementar método de Árvores de Decisão, em específico o algoritmo de classificação J4.8.

A fim de encontrar um padrão, foram realizados 8 conjuntos de testes. O primeiro conjunto abrangeu todas as empresas, enquanto os foram separados em dois conjuntos:

- Um conjunto com 306 empresas para a geração de Árvores de Decisão, ou seja, conjunto de treinamento, em que as empresas estavam classificadas em:
  - 241 adimplentes;
  - 65 inadimplentes.

- Um conjunto com 33 empresas para testar a Árvore gerada, ou seja, conjunto de teste, em que a empresas estavam classificadas em:
  - 25 adimplentes;
  - 8 inadimplentes.

O resultado desse conjunto de dados pode ser visto no quadro abaixo:

Figura 3 – Resultado das Árvores de Decisão

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%

Fonte: LEMOS (2003).

Como é possível perceber, os resultados obtidos do conjunto de treinamento e de teste mostraram que, dentre as Árvores geradas, a do Teste 5 demonstrou ser a com menor taxa de erros no conjunto de treinamento, sendo o melhor padrão. Uma das vantagens encontradas na aplicação dessa técnica consistiu na geração de 40 tipos de regras compreensíveis ao usuário, a partir da Árvore de Decisão gerada pelo Teste 5. Por exemplo, uma das regras consistia na seguinte condição: Se o tempo de conta fosse maior que 25 meses, então a empresa provavelmente seria adimplente.

3.3.2. Aplicação real do Método de Redes Neurais

Nesse estudo, a pesquisadora utilizou o software computacional MATLAB – Neural Networks Toolbox, para implementar método de Redes Neurais, em específico o algoritmo de Retropropagação.

A fim de encontrar um padrão, foram realizados 8 conjuntos de testes. O primeiro conjunto abrangeu todas as empresas, enquanto os foram separados em dois conjuntos:

- Um conjunto com 306 empresas para Treinamento de Redes Neurais, em que as empresas estavam classificadas em:
  - 241 adimplentes;
  - 65 inadimplentes.

- Um conjunto com 33 empresas para testar a Rede gerada, ou seja, conjunto de teste, em que a empresas estavam classificadas em:
  - 25 adimplentes;
  - 8 inadimplentes.

O resultado desse conjunto de dados pode ser visto no quadro abaixo:

Figura 4 – Resultado das Redes Neurais

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%

Fonte: LEMOS (2003).

Como é possível perceber, os resultados obtidos do conjunto de treinamento e de teste mostraram que, dentre as Redes Neurais geradas, a do Teste 6 demonstrou ser a com menor taxa de erros, 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” (LEMOS, 2003).

3.3.3. Resultado do Estudo

Nesse estudo, a pesquisadora apresentou um estudo comparativo que pode ser visto no quadro abaixo:

Figura 5 – Resultado Comparativo

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

Fonte: LEMOS (2003).

Na conclusão apresentada pela pesquisadora em relação à comparação entre as técnicas utilizadas, a técnica de Redes Neurais foi a que apresentou o melhor padrão, por ter a menor taxa de erro. Entretanto, a técnica de Árvores de Decisão levou vantagem no quesito fácil compreensão e entrada de dados. Esta apresentou os resultados com regras detalhadas sobre as informações relevantes na classificação e determinação da sinalização de adimplência ou não.

3.4. Uso de Métodos do Data Mining na Prevenção e Detecção de Fraudes

> "Em apenas um incidente em 1994, um *hacker* russo invadiu o sistema de transferência eletrônica de fundos do Citibank e furtou mais de 10 milhões de dólares, transferindo-os para várias contas espalhadas pelo mundo (Turban; Mc Lean; Wetherbe, 2004, p. 541). Considerando que o Citibank movimenta cerca de um trilhão de dólares por dia, pode-se imaginar como os prejuízos poderiam ter sido maiores caso não fossem tomadas medidas de segurança."
>
> (BASTOS; PEREIRA, 2007)

As fraudes eletrônicas representam atualmente uma grande ameaça tanto às grandes empresas quanto para as pessoas físicas. Com o aumento da facilidade, agilidade e velocidade nas transações comerciais e financeiras cotidianas, devido aos investimentos em tecnologias de informática e de comunicação de dados, todos estão vulneráveis quanto se trata de segurança. O exemplo citado, do Citibank, demonstra que mesmo em grandes instituições, onde normalmente a segurança é uma das maiores preocupações, esta pode ser burlada. Entretanto, a busca pela prevenção e detecção, como formas de proteção necessárias a fim de minimizar perdas e impactos financeiros, tem sido fundamental para a sobrevivência dessas instituições.

As fraudes em sistemas bancários e, principalmente, no que se refere aos cartões de crédito e *internet banking*, têm aumentado gradativamente. Independente da origem da fraude, se de dentro da própria organização ou por pessoas que invadam o sistema, ter ferramentas capazes de perceber tal crime, exige "investimentos para prevenção, detecção e eliminação de fraudes, os quais embora onerem as empresas, reduzindo suas margens de lucro, as protegem de custos muito mais elevados decorrentes de fraudes informatizadas ou eletrônicas" (BASTOS; PEREIRA, 2007).

O controle de fraudes em instituições financeiras, usualmente, é realizado por auditorias internas, as quais utilizam ferramentas baseadas em regras e em análise do comportamento do cliente, capazes de delinear padrões de operações suspeitas.

Dentre as principais estratégias de defesa, a detecção objetiva encontrar uma possível fraude ou operação atípica, visando minimizar os prejuízos ou uma grande fraude. Um dos objetivos da detecção é a correção das falhas de controle, ou seja, caso se encontre alguma falha no sistema que permita uma invasão ou dano, este possa ser corrigido pela equipe.

Em geral o sistema de detecção de fraudes trabalha no acompanhamento de perfis de uso de determinado produto ou serviço, verificando aqueles que se são atípicos. Tendo em vista que as instituições financeiras trabalham com uma quantidade imensa de dados, o uso do Data Mining tem sido essencial na estratégia de detecção de fraudes, já que trabalha com ferramentas inteligentes, tal qual as Redes Neurais capazes de "tentar detectar uma fraude que pode acontecer entre cerca de 50.000 transações" (SANDRI, 2009).

Assim, em casos de fraudes, o Data Mining utiliza associações de fatos históricos de fraudes comparando-as com situações suspeitas, que são identificadas imediatamente quando da ocorrência até posterior confirmação, por apresentarem padrões incomuns.

Segundo BASTOS e PEREIRA (2007):

> "O uso da técnica de *data mining* na detecção de fraudes em cartões de crédito é imprescindível na atualidade. Ao se realizar uma compra pela *internet*, por exemplo, tudo é muito simples para o comprador: "clica" aqui e ali, aguarda uns minutos e compra efetuada. Entretanto, por trás da simplicidade da transação esconde-se todo um sistema de segurança, inclusive quanto à possibilidade de fraudes. O sistema de detecção de fraudes manda alertas, por exemplo, se houver uma compra de diamantes às 4 horas da manhã ou se um usuário fez uso do cartão de crédito em curto intervalo de tempo em local distante um do outro. [...] Uma das ações tomadas pelas administradoras de cartão de crédito e que se encontra atualmente em uso, é a verificação, mediante autorização eletrônica, se o cartão é válido e possui fundos suficientes para que seja efetivada uma compra. Nesse processo avalia-se também se não há registros do cartão ter sido roubado ou "clonado". [... pois] em geral as pessoas costumam fazer uso do cartão de crédito dentro de um padrão bem definido. Ou seja, alguns o usam, por exemplo, somente para compras de mercado, combustível e vestuário. O sistema de redes neurais percebe quando há uso do cartão em padrões diferentes, neste caso exemplificado, em gastos excessivos com lazer e cultura."

Como visto no exemplo acima, cada cliente tem um perfil e, no caso de padrões de transações ilícitas, estas são automaticamente detectadas por sistemas de Redes Neurais que são treinados para reconhecer estes tipos de situações.

Dentre as técnicas aplicadas na detecção de fraudes em cartões de crédito, a mais utilizada é a de Redes Neurais, pois segundo TREZUB (2004), "utiliza técnicas de inteligência artificial para aprender novos padrões de fraude e conter novas tentativas de fraude sem intervenção humana". Esse fato é relevante já que os sistemas convencionais não possuem a capacidade de se adaptar aos golpes e às fraudes que vão se aprimorando e modificando-se diariamente.

Na área bancária, o método de Redes Neurais funciona da seguinte forma:

> “Softwares que utilizam redes neurais são integrados com os sistemas bancários de gerência de cartão e sistemas de autorização. Estes tipos de softwares consistem em uma gama de programas que reconhecem padrões a partir do comportamento do portador do cartão, o qual é fruto de uma longa história em trabalhos de reconhecimento de padrões utilizando computadores. O sistema monitora o comportamento do portador do cartão procurando por volumes transacionais, quantias e localidades incomuns, conforme os hábitos do titular. Monitora também os tipos de comerciantes que são utilizados e padrões que não combinam com o histórico de utilização do cartão. O sistema contabiliza pontos para cada transação, dando valores maiores para aqueles em que ele suspeita que sejam fraudulentos. O sistema pode monitorar os resultados em tempo real ou analisar os históricos periodicamente para reforçar as próximas detecções da rede neural. “(SANDRI, 2009)

Mais uma vez, vale salientar que o uso dos sistemas de Redes Neurais associados aos sistemas bancários não deve desconsiderar os profissionais compõem a equipe de prevenção, pois, apesar do método de Redes Neurais ser responsável pela detecção de fraudes, a tomada de decisão fica a cargo dos profissionais responsáveis.

Assim, o uso do método de Redes Neurais é uma das principais ferramentas responsáveis pela crescente diminuição de fraudes bancárias pela habilidade de detectar comportamentos fraudulentos por meio da análise das transações comparadas. Ou seja, o método de Redes Neurais realiza a comparação entre os padrões de fraude e tipos e fraudes existentes ainda desconhecidas. Segundo SANDRI (2009), “através de Redes Neurais: para cada novo caso de fraude, o sistema calcula um valor de pontuação conforme sua similaridade com um padrão conhecido. Atualmente é a técnica mais utilizada, pois oferece melhores resultados”.

As Redes Neurais, entretanto, têm a necessidade de utilizar dados históricos recentes de pelo menos seis meses de atividades transacionais para compor seus conjuntos de treinamento e teste, gerando assim bons padrões de aprendizagem de fraudes.

CONCLUSÃO

Após o estudo e análise efetuados para conclusão desse trabalho ficou claro qual a importância dos métodos do Data Mining na determinação de padrões de comportamentos, auxiliando no processo decisório.

O uso dos métodos do Data Mining, associados aos processos do KDD – *Knowledge Discovery in Databases*, tais como Seleção de Dados, Limpeza de Dados, Transformação de Dados e Interpretação e Avaliação de Dados, permite às organizações trabalhar com informações implícitas e a partir dessas buscar por padrões de consumo e tendências de mercado, bem como mudar estratégias adotadas, a fim de gerar lucro significativo ou minimizar perdas.

Dentre os métodos de Data Mining abordados neste trabalho, ao se verificar sua aplicabilidade, em específico em instituições bancárias e de crédito que trabalham com imensas bases de dados e com transações financeiras com volumes elevados e de alto risco, destacaram-se a Descoberta de Regras por Associação, Árvores de Decisão e Redes Neurais por apresentarem resultados eficazes e modelos e padrões adequados aos tipos de dados utilizados por essas organizações, principalmente em práticas como análise de crédito, fidelização de clientes, vendas de pacotes e, finalmente, na detecção de fraudes.

Por fim, vale salientar que, apesar da relevância do Data Mining e de seus benefícios, o papel do analista, gestores e pesquisadores não deve ser desconsiderado, mesmo com o uso das melhores técnicas de Data Mining, já que a aplicação e a análise adequada desses dependem de bons profissionais.

REFERÊNCIAS BIBLIOGRÁFICAS E ACESSOS

ADAMO, Jean-Marc. **Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms**. New York: Springer, 2000.

BASTOS, Paulo S. S.; PEREIRA, Roberto M. **Fraudes Eletrônicas: O que há de novo?** Rio de Janeira: Revista de Contabilidade do Mestrado em Ciências Contábeis da UERJ, v. 12, n.2, p.1, maio/agosto, 2007.

CARVALHO, Juliano V. **Reconhecimento de Caracteres Manuscritos utilizando Regras de Associação.** Campina Grande: Dissertação de Mestrado da Universidade Federal da Paraíba, 2000.

CASTANHEIRA, L. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões.** Belo Horizonte: Dissertação de Mestrado da Universidade Federal de Minas Gerais, 2008.

DZEROSKI, Saso; LAVRAC, Nada. **Relational Data Mining**. Heildelberg: Sprienger-Verlag, 2001.

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques.** San Francisco: Elsevier, 2006.

LEMOS, Elaine P. **Análise de Crédito Bancário com o uso de Data Mining: Redes Neurais e Árvores de Decisão.** Curitiba: Dissertação de Mestrado da Universidade Federal do Paraná, 2003.

MODERNO **Dicionário da Língua Portuguesa**. Disponível em: <http://michaelis.uol.com.br/moderno/portugues/index.php>. Acesso em 23 out. 2011.

NANGIYALIL, Sajeev G. **Estudo de Ferramenta de KDD & Mineração de Dados**. In: Trabalho Monográfico. São Paulo. UNICID-SP, 2007.

NAVEGA, Sergio. **Princípios Essenciais do Data Mining**. São Paulo: Anais do Infoimagem 2002, Cenadem, Novembro. Disponível em < http://www.inteliwise.com/reports/i2002.pdf >. Acesso em: 24 agosto 2011.

PINTO, Rodrigo A. **Business Intelligence (BI) – A inteligência influenciando os negócios.** In: Trabalho Monográfico. São Paulo. FATEC-SP, 2004.

ROKACH, Lior; MAIMON, Oded. **Data Mining with Decision Trees: Theory and Applications**. London: World Scientific, 2008.

SANDRI, André. **Detecção de Fraude e Prevenção Utilizando Inteligência Artificial.** Disponível em < http://andresandri.com.br/artigos/artigos.html >. Acesso em 15 out. 2011.

SCAFF, Vinícius P.; LIMA, Renato da S.; ALMEIDA, Dagoberto A. **Sistemas de Informação como ferramenta de Apoio à Decisão.** Bauru: XI SIMPEP, 2005. Disponível em < http:// simpep.feb.unesp.br >. Acesso em 05 set. 2011.

SETZER, Valdemar W.; SILVA, Flávio S. C. **Banco de Dados: Aprenda o que são, melhore seu conhecimento, construa os seus.** São Paulo: Edgard Blucher, 2005.

SFERRA, Heloísa H.; CORREA, Ângela C.J. **Conceitos e Aplicações de Data Mining**. Revista de Ciência e Tecnologia. V.11 Nº 22. Disponível em: <http://www.unifra.br/professores/EDUARDO/Artigo%208.pdf>

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining – Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

THOMÉ, Antonio C. G. **Redes Neurais – Uma ferramenta para KDD e Data Mining**. Disponível em < http://equipe.nce.ufrj.br/thome/grad/nn/mat.../apostila_**kdd**_mbi.pdf >. Acesso em 15 out. 2011.

VENTURA, R. **Bancos tendem a aperfeiçoar o Data Mining.** Disponível em: <http://webinsider.uol.com.br/2003/06/22/bancos-tendem-a-aperfeicoar-o-data-mining/>. Acesso em: 02 setembro 2011.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining: Practical Machine Learning Tools and Techniques**. San Francisco: Elsevier, 2005.