Sonia Kaoru Shiba

Desenvolvimento de Modelo de Processo de Extração de Conhecimento em
Banco de Dados para Sistemas de Suporte à Decisão

São Paulo
2008

Sonia Kaoru Shiba

Desenvolvimento de Modelo de Processo de Extração de Conhecimento em
Banco de Dados para Sistemas de Suporte à Decisão

Dissertação apresentada à Escola Politécnica
da Universidade de São Paulo para obtenção
do título de Mestre em Engenharia Elétrica.

Área de Concentração:
Sistemas Eletrônicos

Orientador:
Prof. Dr. Francisco Javier Ramirez Fernandez

São Paulo
2008

Shiba, Sonia Kaoru
Modelagem de processo de extração de conhecimento em
banco de dados para sistemas de suporte à decisão / S.K. Shiba.
-- São Paulo, 2008.
p.103

Dissertação (Mestrado) - Escola Politécnica da Universidade
de São Paulo. Departamento de Engenharia de Sistemas Eletrô-
nicos.

1.Banco de dados orientado a objetos 2.Conhecimento (Mo-
delagem) 3.Teoria da decisão (Inferência estatística) I.Univer-
sidade de São Paulo. Escola Politécnica. Departamento de
Engenharia de Sistemas Eletrônicos II.t.

DEDICATÓRIA

Dedico este trabalho a meus pais.

Agradecimentos

Este trabalho representa a superação de um grande desafio pessoal, uma conquista que me proporcionou grandes desafios e que alcancei graças ao apoio de meu orientador, Prof. Francisco Javier Ramirez Fernandes, que muito contribuiu para o meu crescimento científico.

Agradeço aos amigos que acreditaram em meu potencial para o ingresso ao programa de mestrado desta instituição, os professores Roberto Almeida e João Viane, aos amigos Luis Carlos Garcia, Sergio Donizeti e Eduardo Maia, pela colaboração nas minhas pesquisas a partir das discussões sobre Business Intelligence, Data Warehouse e ferramentas ETL. Aos amigos Roberto Teruo Matsuda e Daniele Rodrigues, pelas primeiras discussões que tivemos sobre sistemas de suporte à decisão.

Dedico um agradecimento especial ao Prof. João Francisco Justo Filho, que me auxiliou nas revisões deste trabalho, pelo incentivo constante nos momentos mais difíceis e por alguns momentos, ser pai e mãe do nosso filho João Victor. Ao meu filho João Victor, por me amar e acreditar que sou quase tudo em sua vida, mesmo tendo que dividir seu espaço com minhas pesquisas.

RESUMO

Este trabalho apresenta a modelagem de um processo de extração de conhecimento, onde a aquisição de informações para a análise de dados têm como origem os bancos de dados transacionais e *data warehouse*. A mineração de dados focou-se na geração de modelos descritivos a partir de técnicas de classificação baseada no Teorema de Bayes e no método direto de extração de regras de classificação, definindo uma metodologia para a geração de modelos de aprendizagem. Foi implementado um processo de extração de conhecimento para a geração de modelos de aprendizagem para suporte à decisão, aplicando técnicas de mineração de dados para modelos descritivos e geração de regras de classificação. Explorou-se a possibilidade de transformar os modelos de aprendizagem em bases de conhecimento utilizando um banco de dados relacional, disponível para acesso via sistema especialista, para a realização de novas classificações de registros, ou então possibilitar a visualização dos resultados a partir de planilhas eletrônicas. No cenário descrito neste trabalho, a organização dos procedimentos da etapa de pré-processamento permitiu que a extração de atributos adicionais ou transformação de dados fosse realizada de forma iterativa, sem a necessidade de implementação de novos programas de extração de dados. Desta forma, foram definidas todas as atividades essenciais do pré-processamento e a seqüência em que estas devem ser realizadas, além de possibilitar a repetição dos procedimentos sem perdas das unidades codificadas para o processo de extração de dados. Um modelo de processo de extração de conhecimento iterativo e quantificável, em termos das etapas e procedimentos, foi configurado vislumbrando um produto final com o projeto da base de conhecimento para ações de retenção de clientes e regras para ações específicas com segmentos de clientes.

ABSTRACT

This work presents a model of knowledge discovery in databases, where the information for data analysis comes from a repository of transactional information systems and data-warehouse. The data mining focused on the generation of descriptive models by means of classification techniques based on the Bayes' theorem and a extraction method of classification rules, defining a methodology to propose new learning models. The process of knowledge extraction was implemented for the generation of learning models for support the make decision, applying data mining for descriptive models and generation of classification rules. This work explored the possibility of transforming the learning models in knowledge database using a relational database, to be accessible by a specialist system, to classify new records or to allow the visualization of the results through electronic tables. The organization of the procedures in the pre-processing allowed to extract additional attributes or to transform information in an interactive process, with no need of new programs to extract the information. This way, all the essential activities of the pre-processing were defined and the sequence in which these should be developed. Additionally, this allowed the repetition of the procedures with no loss of units for the process of information extraction. A model of process for the interactive and quantifiable extraction of knowledge, in terms of the stages and procedures, was idealized in order to develop a product with the project of the knowledge databases for actions of retention of clients and rules for specific actions within clients' segments.

Índice

1. Introdução	9
1.1 - Extração de Conhecimento a Partir de Bases de Dados	11
1.2 - Ferramentas de Suporte à Decisão: o Processamento Transacional e os Repositórios Históricos	13
1.3 - O ambiente data warehouse nas corporações	14
1.4 - Sistemas especialistas	16
2. Extração de Conhecimento	20
2.1- Elementos do Processo KDD	23
2.2 - Definição do Processo de Extração de Conhecimento em Base de Dados	25
2.3 - Modelos de Processo KDD.....	27
2.3.1 - Modelo de Fayad, Piatetsky-Shapiro e Smyth.....	28
2.3.2 - Modelo de Williams e Huang.....	30
2.2.3 - Modelo de Hipp, Güntzer e Nakhaeizadeh.....	31
2.4 - Organização do Processo de Extração de Conhecimento.....	32
2.5 - Ferramentas de suporte ao processo KDD	34
2.6 - Pré-Processamento de Dados.....	36
2.7 - Seleção de Dados	42
3. Mineração de Dados	52
Introdução	52
3.1 - Conceitos de Mineração de Dados	53
3.1.1- Caracterização e discriminação de classes e conceitos	54
3.1.2 - Padrões por frequência, associações e correlações entre dados	55
3.1.3 - Análise de agrupamentos de dados ou clustering	56
3.1.4 - Classificação e predição de eventos	56
3.2 - Modelagem Preditiva: Técnicas de Classificação	58
3.4 - Modelos de Classificação Baseados em Regras	61
3.5 - Inferência Probabilística para Tratamento de Incertezas em Modelos de Classificação	66
3.6 - Classificadores Bayesianos – Naïve Bayes	68
3.6.1 - Classificação de eventos utilizando o Teorema de Bayes	68
3.7 - Pós- Processamento do Processo KDD	69
4. Aplicações do Processo KDD	72
4.1 - Estudo de Retenção de Clientes.....	73
4.1.1 - Metodologia aplicada ao Pré-Processamento de dados.....	74
4.1.2 - Modelo de retenção com subclasse fixa	80
4.1.3 - Modelo de classificação descritivo para estudo de retenção	91
4.2 - Aplicação dos modelos descritivos em estudos de segmentos de clientes	94
4.3 - Otimização do processo KDD	98
5. Conclusões	101
Referências Bibliográficas	106

1. Introdução

Nos últimos anos, instituições de pesquisa, governamentais e empresas de diversos ramos de atuação vêm investindo fortemente em novas abordagens aplicáveis aos processos decisórios. Uma mudança cultural na composição dos elementos essenciais para o processo de decisão têm orientado a busca e a utilização de ferramentas de apoio, bem como a implementação de serviços para a aquisição da informação. Além disso, a automatização das atividades organizacionais por meio de *softwares* e a diminuição dos custos de armazenamento de dados contribuíram para obtenção de informação confiável de forma rápida e segura (Berry & Linoff, 2004).

Uma vez garantida a aquisição e armazenamento da informação, os procedimentos de utilização dos dados vem sendo revistos para responder a questionamentos comuns no ambiente corporativo, tais como: identificação rápida e precisa de novos comportamentos nos negócios; composição e

evolução de novos produtos e serviços; construção de modelos para a predição de eventos, dentre outros (Kurgan & Musilek, 2006).

A disponibilidade de extensos bancos de informação contribuiu para que os processos decisórios oriundos de práticas com pouco embasamento conceitual fossem praticamente abandonadas nas corporações. Por outro lado, a expansão da quantidade de informação nos bancos de dados passou a figurar como um elemento dificultador dos processos. Neste contexto, o volume de registros envolvidos na análise de decisão fez com que a abordagem de análise tradicional, por meio de planilhas e relatórios, um procedimento caro e demorado, sendo no final pouco elucidativo para a obtenção de informação nova e relevante no comportamento dos dados.

Este trabalho elabora uma proposta de extração de conhecimento em bases de dados, tratando seus elementos essenciais:

- A informação, que é tratada por processos e armazenada em banco de dados;
- As técnicas de extração de padrões e as operações necessárias para o tratamento de dados;
- A visualização dos resultados e a validação dos padrões identificados;
- A disponibilização do conhecimento para sistemas de apoio à decisão.

Este trabalho tem como principal objetivo desenvolver um processo de Extração de Conhecimento em Base de Dados de tal forma a construir bases de conhecimento representativas de áreas de domínio, que possam ser utilizadas por Sistemas Especialistas de Apoio à Decisão.

O processo de Extração de Conhecimento em Base de Dados é discutido no capítulo 2, com a apresentação dos modelos de processos idealizados por Fayad, Piatetsky-Shapiro e Smyth (Fayad, Piatetsky-Shapiro & Smyth, 1996), Williams e Huang (Williams & Huang, 1996) e Hipp, Güntzer e Nakhaeizadeh (Hipp, Güntzer & Nakhaeizadeh, 2002). O capítulo apresenta a

definição de um processo com base nos modelos apresentados e o detalhamento do fluxo de trabalho em cada etapa do processo de extração, com os participantes e os recursos necessários, desde a captação dos dados para análise, a extração dos padrões, até a sua disponibilização em bases de conhecimento para sistemas especialistas.

O capítulo 3 discute as principais técnicas e tarefas de mineração de dados, a aplicação dessas técnicas em sistemas reais e os resultados de cada técnica apresentada. O capítulo 4 apresenta problemas reais onde se aplica o modelo de extração de conhecimento proposto, com a exploração de dois cenários muito comuns em áreas de planejamento de vendas: a retenção e a aquisição de clientes. A partir destes dois problemas foi proposto um processo de extração de conhecimento dividido em três etapas, para otimizar as atividades de criação de modelos de classificação descritivos e modelos de classificação segmentados.

1.1 - Extração de Conhecimento a Partir de Bases de Dados

Na atualidade, o processo de transformação dos dados pertencentes a uma determinada base em conhecimento útil, em sua maior parte, é resultado de uma análise tradicional, por meio de relatórios ou ferramentas específicas. Esses recursos permitem a visibilidade de um conjunto de informações cuja avaliação é executada por analistas com grande conhecimento nas suas áreas de atuação. Em termos práticos, isso significa delegar a tarefa de elucidar comportamentos e tendências refletidas nos dados a um grupo restrito de indivíduos especialistas, os analistas de negócio, que avaliam informações obtidas a partir de alguma ferramenta ou software com o intuito de identificar elementos significativos para uma futura tomada de decisão (Fayad, Piatetsky-Shapiro & Smyth, 1996).

Essa prática de análise de dados foi, por muitos anos, o único procedimento utilizado para a extração de conhecimento, empregado e aceito para as tomadas de decisão estratégicas dentro de uma empresa. Entretanto, esses procedimentos vem se revelando cada vez mais onerosos e demorados,

além de gerar, muitas vezes, resultados extremamente subjetivos e imprecisos (Yoon & Kerschberg, 1993). A identificação de um padrão, ou seja, de uma informação útil, tornou-se uma tarefa exaustiva e difícil de ser executada por um especialista, devido à necessidade de se analisar um conjunto cada vez maior de dados dentro de prazos cada vez mais restritos. Além disso, mesmo que as corporações ainda se utilizem desses procedimentos para a extração de conhecimento, esta estratégia não garante que o comportamento dinâmico do negócio está sendo considerado nos processos de seleção de dados.

Com a grande oferta de recursos para a captação e armazenamento de dados, tornou-se comum encontrar sistemas de bancos de dados capazes de gerenciar milhões de transações diárias, como ocorre, por exemplo com o banco de dados criado para a empresa americana Wal-Mart (Babcock, 1994). A empresa Mobil Oil Corporation desenvolveu um *data warehouse* capaz de armazenar mais de 10 terabytes de dados relacionados à exploração de petróleo (Harrison, 1998). Num contexto acadêmico, a base de dados para o projeto do Genoma Humano foi preparada para coletar vários gigabytes de informação sobre o código genético humano (Fasman, Cuticchia & Kingsbury 1994). Diante desse cenário, a necessidade de se desenvolver técnicas e ferramentas com a capacidade de extração de conhecimento de forma inteligente e automatizada levou ao surgimento de um campo de pesquisa conhecido por *Knowledge Discovery in Databases*, abreviado por KDD (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

Este trabalho apresenta uma proposta para formalizar o processo de extração de conhecimento em corporações, onde a complexidade dos processos de negócio e a grande de massa de dados armazenados tornou crítica a tarefa de aquisição de conhecimento. O processo de extração de conhecimento tem como propósito a apresentação de um modelo aplicável a qualquer domínio de conhecimento.

1.2 - Ferramentas de Suporte à Decisão: o Processamento Transacional e os Repositórios Históricos

Uma empresa de pequeno porte¹ em relação à quantidade de transações, pode planejar as ações de relacionamento com seus clientes a partir da percepção das suas necessidades, preferências e interações ocorridas no passado. Já uma empresa de grande porte, para configurar um plano semelhante, depende da captura e centralização das informações, já que mantém uma rede de relacionamento maior, com vários canais de interação com o cliente sendo utilizados de forma concorrente e maior quantidade de profissionais envolvidos na gestão das informações. Um dos desafios, para as empresas que operam nessas condições, é explorar alternativas para perceber e aprender a partir das interações com os seus clientes, de modo a perceber seus hábitos e preferências de consumo.

Através da aplicação de recursos da tecnologia da informação, empresas de grande porte podem criar ambientes que centralizam informações sobre as interações que ocorrem com seus clientes, cuja origem podem ser: servidores de aplicações web, dados de caixa eletrônicos, scanners de pontos de venda, sistemas de automatização para call-centers. A integração de sistemas que atedem a diversas áreas de uma empresa, permitindo a interação entre processos organizacionais e processamento de informações é conhecido como ERP – *Enterprise Resource Planning*. Todas essas fontes de dados podem ser úteis para a mineração de dados.

Quando um cliente entra em contato com uma grande loja para se informar sobre detalhes de algum produto que deseje comprar, um registro da ligação é armazenado, constando, entre outras informações, o tempo de duração da ligação, o número discado, a data do contato, os setores e atendentes para os quais a ligação foi direcionada e um código identificador do cliente. Da mesma forma, inúmeras ligações com os mesmos objetivos estão

¹ Em relação ao porte das empresas, qualifica-se neste trabalho a quantidade de transações em banco de dados que é realizada diariamente, o qual apresenta variações significativas de acordo com o ramo de atuação: empresas prestadoras de serviço e varejistas processam uma quantidade de dados.

sendo registrados vindos de diversos locais. Se esses contatos concretizam alguma venda, a combinação de outras informações pessoais e a confirmação do pagamento, por exemplo a partir de sistemas que validam dados do cartão de crédito e aprovam o valor requisitado, há uma transação de venda, e um registro de venda será criado no banco de dados.

Esses registros de dados não são gerados com o objetivo de se realizar uma mineração de dados, eles surgem de necessidades operacionais de uma empresa, porém eles contém informações valiosas sobre os clientes e podem ser utilizados numa mineração de dados. Empresas de telefonia utilizam detalhes de registros de chamadas para descobrir números de telefones residenciais cujas chamadas apresentam os mesmos padrões de uma empresa, para que possam oferecer a esses clientes alguns serviços especiais para pessoas que operam seus negócios a partir de suas residências. Empresas utilizam históricos de vendas para envio de catálogos a seus clientes, com base nos produtos mais vendidos e empresas com site de vendas utilizam o histórico das vendas anteriores para determinar os produtos que serão exibidos aos clientes quando estes retornam ao site.

Os sistemas de processamento transacional representam o ponto inicial de entrada das informações de uma empresa, e são sensíveis ao dinamismo do mercado consumidor. Por essa razão, eles têm sido o foco de investimentos para que sejam modelados para a captura das informações relevantes, minimizando perdas de informações, bem como a fácil recuperação e atualização dos dados quando necessário.

1.3 - O ambiente data warehouse nas corporações

Uma empresa com foco no cliente considera cada registro de interação com o cliente como perspectiva de negócio. Todo contato do cliente com um *call-center*, toda operação registrada em pontos de venda no varejo, todo catálogo enviado, ou ainda uma visita do cliente numa página web da empresa devem ser vistas como oportunidades de aprendizado. Porém, o

aprendizado não se restringe simplesmente em reunir dados, pois muitas empresas podem reunir centenas de gigabytes de dados e tal fato não garante que essas empresas detém o conhecimento sobre as preferências de seus clientes. Para essas empresas, os dados são coletados porque são necessários para atender a um grupo de atividades operacionais, tal como controle de estoque, movimentações e faturamento, e uma vez que os dados atenderam a essas necessidades, são armazenados em disco ou fita, ou então descartados.

Para que o aprendizado ocorra, uma preparação no ambiente de dados deve ser formalizada, reunindo dados de diversas fontes em um único repositório. Desta forma, todas as interações com os clientes, sejam vendas, contatos telefônicos e processos, entre outros, devem ser reunidos e organizados num modelo coerente. Essa abordagem de manutenção dos dados históricos é conhecida como *data warehouse*.

Um dos aspectos mais importantes do *data warehouse* é a capacidade de acompanhar as relações com os clientes ao longo do tempo. Muitos padrões de interesse só se tornam visíveis quando apuradas dentro de um determinado período, por essa razão, um dos requisitos para os projetos de mineração de dados é a manutenção de um repositório de dados como o *data warehouse*. Com os dados de um *data warehouse*, os analistas conseguem identificar a frequência com que um determinado produto é comprado, quais promoções os clientes costumam responder, quais canais de venda de melhor desempenho, dentre outros.

Um *data warehouse* apropriado deve fornecer acesso aos dados em um formato mais amigável que o acesso aos dados dos sistemas transacionais. Ele deve reunir dados de diversas fontes de informação, e mantendo-os limpos, transformados e vinculados. Embora alguns repositórios não alcancem um padrão de qualidade das informações, ainda assim, continuam sendo fontes de dados importantes de uma corporação para a gestão das relações com seus clientes.

1.4 - Sistemas especialistas

Nos últimos anos, a necessidade de ferramentas de apoio nos processos decisórios de uma organização impulsionou o desenvolvimento dos chamados sistemas especialistas (SE) baseados em um modelo organizacional que destaca a gestão do conhecimento e da informação. De fato, a arquitetura ideal para os sistemas especialistas baseia-se no desenvolvimento de componentes de software capazes de representar os elementos essenciais de um processo de decisão: a informação e o critério de decisão (Yoon & Kerschberg, 1993). De acordo com as definições mais recentes, pode-se definir os sistemas especialistas como sistemas que atuam sobre um grupo de informações sob determinados critérios de decisão para a resolução de problemas, exibindo comportamento e desempenho semelhante a um especialista humano (Luger, 2004).

Os sistemas especialistas atuam sobre um conjunto específico de problemas, ou seja, uma área de domínio. Por exemplo, um sistema especialista de suporte a diagnósticos médicos será eficaz por inferir com certa habilidade nas avaliações clínicas, e assim como um médico, a amplitude de suas inferências dependerá do histórico de dados disponível ou seja, das experiências registradas em alguma base de conhecimento, como resultado da combinação de elementos teóricos para entendimento de um problema com regras heurísticas para a sua resolução (Luger, 2004).

Atualmente, observam-se propostas de sistemas especialistas para a resolução de problemas de uma ampla escala de domínios, tais como medicina, matemática, engenharias, química, geologia, ciência da computação, economia, direito e educação. Para facilitar o entendimento sobre os tipos de sistemas especialistas, Waterman (Waterman, 1986) propõe categorias específicas na classificação dos SE's:

- Interpretação: oferecer conclusões a partir de dados brutos;
- Predição: avaliar possíveis consequências a partir de situações conhecidas;

- Diagnóstico: determinar a causa em situações e combinações complexas de sintomas observados;
- Projeto: propor configurações de sistemas que alcancem objetivos de desempenho, e simultaneamente tratem as restrições de projeto;
- Planejamento: estabelecem um conjunto de ações para alcançar um conjunto de objetivos, dadas as condições iniciais e restrições de tempo de execução;
- Monitoramento: comparar os comportamentos observados aos comportamentos esperados;
- Controle: governar o comportamento de um ambiente complexo;
- Instrução: dar assistência ao processo de aprendizagem em domínios técnicos.

Uma vez determinada a categoria do sistema especialista, o projeto de desenvolvimento envolve a formação de uma base de conhecimento a partir de um banco de dados relacionado ao problema, e a definição de uma estratégia para a codificação dos critérios de decisão e uma vez codificados, irá compor a máquina de inferência do sistema especialista.

Os sistemas especialistas - SE de primeira geração foram construídos sob uma ótica unificada dos elementos de suporte a decisão: informação e critérios de decisão eram codificados na estrutura da máquina de inferência. Posteriormente, esse padrão de arquitetura dos sistemas especialistas mostrou-se pouco flexível às necessidades dos usuários, uma vez que o esforço necessário para a alteração ou incorporação de novas regras era uma tarefa complexa, que demandava um custo alto de manutenção do software, e muitas vezes fora de prazos aceitáveis (Schreiber *et al.*, 1999), (Abdullah *et al.*, 2004).

Uma mudança estratégica no desenvolvimento de software, com a utilização dos conceitos de orientação a objetos e visualização do comportamento sistêmico por meio de modelos, orientou o desenvolvimento de sistemas especialistas a partir da representação de seus elementos básicos: a *máquina de inferência* e a *base de conhecimento*, conforme a figura 1.1. Essa

nova abordagem, voltada para a componentização dos elementos do sistema, possibilita que novos conceitos do domínio do conhecimento, sejam novos critérios ou regras, possam ser incorporados ou reutilizados em outros problemas do mesmo domínio (Schreiber *et al.*, 1999), (Speel *et al.*, 2001).

Fig. 1.1: Arquitetura de um Sistema Especialista - SE.

A máquina de inferência passou a ser modelada como um pacote de software, cujos componentes são implementações de algoritmos de suporte à decisão e a base de conhecimento, um repositório de dados para a realização de inferências dentro dos limites do domínio do conhecimento (Luger, 2004).

Outras características que interferem no projeto dos sistemas especialistas são definidos pelo próprio usuário:

- ✓ Tempo de resposta compatível com a necessidade do usuário. Os sistemas de suporte à decisão necessitam de uma arquitetura flexível em relação ao acesso à base de conhecimento, prevendo acessos remotos e concorrentes em ambientes críticos. Necessitam gerenciar o volume de informações – sejam regras ou outros indicadores, que serão manipuladas por algoritmos de suporte à decisão.

- ✓ A interface homem-máquina deve ser amigável em relação às funções de suporte à decisão, prevendo situações em que o usuário irá utilizar de diferentes combinações de atributos para avaliação de um registro único ou um lote de registros. O modelo de um sistema de suporte às atividades operacionais é conceitualmente diferente de um sistema de suporte à decisão, o qual poderá exibir diferentes resultados de análise para um mesmo conjunto de dados, de acordo com o tipo de algoritmo de suporte à decisão que foi adotado. Essa diferença conceitual deve ser conhecida por seus usuários.
- ✓ Componentes de suporte à decisão podem ser embutidos em sistemas de suporte às atividades organizacionais: nestes casos, um componente de software pode ser agregado a um sistema para filtrar objetos ou compor grupos de especialidades diferentes. O componente pode implementar uma ou mais regras de classificação e atuar sobre um conjunto de dados, acessando uma base de conhecimento, sendo que essas ações ficam transparentes ao usuário final.

Em relação ao processo de desenvolvimento de sistemas especialistas, Luger faz um breve resumo, apontando a utilização de processos e metodologias de desenvolvimento de softwares alternativas ao modelo clássico (Luger, 2004): “... a construção de sistemas especialistas requer um ciclo de desenvolvimento não tradicional baseado em prototipação o mais cedo possível e em revisão incremental de código”. Em relação aos processos de extração de conhecimento em bancos de dados, têm-se a proposta de um processo cíclico, organizado sistematicamente, com possibilidades de repetição e que mantém procedimentos de avaliação contínua de desempenho dos modelos.

2. Extração de Conhecimento

Nos últimos anos, observou-se um rápido crescimento no volume de registros armazenados em banco de dados (BD). Estima-se que alguns bancos de dados já tenham atingido uma quantidade de registros em torno de 10^9 objetos, exemplo que pode ser observado no banco de dados do projeto Genoma Humano (Fasman, Cuticchia & Kingsbury 1994). Tal aspecto em termos de volume de dados se deve, em grande parte, à extensa oferta de recursos com capacidade de armazenamento cada vez maior e o desenvolvimento de softwares para suporte às atividades organizacionais nas empresas (Fayyad,

Piatetsky-Shapiro & Smyth, 1996), (Dunham, 2002),(Han & Kamber, 2001),(Hand *et al.*, 2001).

A alta disponibilidade de recursos para armazenamento de dados também permitiu às organizações um aumento significativo nos investimentos para a capacitação de seus ambientes no que se refere à captura, transformação e retenção de informações, dotando-os de softwares funcionalmente capazes de suportar todo o fluxo das transações de negócio (Berry & Linoff, 2004). Por outro lado, a alta disponibilidade desses recursos seria apenas um dos elementos essenciais para um propósito maior: a transformação desses dados coletados em conhecimento. Para a aquisição de conhecimento a partir dos dados, seria necessário que essas informações, uma vez capturadas, fossem trabalhadas por métodos que permitissem a identificação de padrões.

Durante muitas décadas no cenário corporativo, o processo de reconhecimento de informações úteis para a tomada de decisão foi resultado de uma análise tradicional, baseada na geração e análise de extensos relatórios ou utilização de ferramentas específicas, tais como as planilhas eletrônicas (McGarry, 2005). Esse tipo de análise foi essencial para os processos de tomada de decisão e por muitos anos, sendo que essas ferramentas representaram por muitos anos, uma das únicas alternativas disponíveis. Por outro lado, com o crescimento das bases de dados e conseqüentemente a quantidade de informações envolvidas, tal metodologia se tornou inadequada para oferecer os mesmos resultados aos seus usuários finais. Em termos práticos, tal metodologia significava delegar a um grupo restrito de indivíduos especialistas a tarefa de elucidar os comportamentos e tendências refletidas nos dados. Esses profissionais, os analistas de negócio, avaliavam as informações obtidas a partir de alguma ferramenta ou software com o intuito de identificar elementos significativos para uma futura tomada de decisão (Fayyad, Piatetsky-Shapiro & Smyth, 1996). A atividade de correlacionar informações depende de um bom conhecimento acerca dos negócios, com a finalidade de avaliar a relevância da informação identificada (McGarry, 2005). Tal tarefa pode ser extremamente difícil

de ser executada com o máximo desempenho na ausência de uma metodologia e ferramentas de suporte adequadas.

O processo de extração de conhecimento a partir de bases de dados, também conhecido como *Knowledge Discovery in Databases* (KDD), surgiu com o objetivo de oferecer suporte em diversas áreas do conhecimento (Hand *et al.*, 2001), reunindo metodologias e ferramentas de aprendizagem para cenários onde é comum a manipulação de volumes consideráveis de dados. O processo KDD consiste num conjunto de etapas ou fases, orientadas por atividades direcionadas para a extração e manipulação de registros e aplicação de técnicas de mineração de dados, finalizadas por uma avaliação dos padrões extraídos, tal como representado na figura 2.1. Para as corporações, a implementação de processos KDD depende de investimentos consideráveis em profissionais e ferramentas, além de preparar um ambiente independente de banco de dados históricos para a aplicação dos procedimentos de mineração de dados.

As etapas onde ocorrem a aquisição e a convergência para uma base de informações históricas requerem investimentos em ferramentas ETL (“*extract, transform and load*”) e implementação de programas específicos, modelagem da estrutura de data warehouse, implementação e manutenção dos programas de extração de dados. O processamento da extração deve ocorrer sem concorrência com os programas de produção – processamento transacional (Berry & Linoff, 2004), (Teorey, Lightstone & Nadeau, 2006).

Uma vez implementado, o processo KDD torna-se um recurso valioso para as decisões estratégicas de uma empresa, podendo interferir diretamente na composição de novos serviços e desenvolvimento de produtos, redefinição em campanhas de vendas e divulgação de produtos e análise da evolução da carteira de clientes, análise e gestão de riscos, dentre outros (Berry & Linoff, 2004), (Parr Rud, 2001), (Pyle, 2003).

Fig. 2.1: Processo KDD: ciclo de vida do processo de extração de conhecimento.

2.1- Elementos do Processo KDD

Do ponto de vista de organização de processos, o processo de extração de conhecimento pode ser implementado sob dois aspectos diferentes: uma delas envolve a definição dos conceitos e a formação de regras aplicáveis à área de domínio a partir dos registros de experiências de profissionais altamente capacitados na resolução de problemas da área de domínio (Luger, 2004). Em contrapartida, a outra abordagem envolve um processo de coleta e análise de informações, sendo estas originárias de um banco de dados. Ambos processos dependem da interação de elementos essenciais na gestão de conhecimento de uma área de domínio – tais como pessoas, ambiente de dados e algum processo de revisão das informações (Kurgan & Musilek, 2006). O diferencial do processo KDD está na incorporação de metodologias aplicadas à análise de

dados, com o uso de ferramentas estatísticas e técnicas de mineração de dados os quais permitem ampliar o espaço de dados na extração de conhecimento. A análise de vários modelos de extração de conhecimento, permite vislumbrar um modelo básico, determinado por elementos essenciais para todo processo em que se destina a extração de conhecimento em bancos de dados, conforme exibido na figura 2.2.

Fig. 2.2- Elementos essenciais do processo KDD.

Essa visão dos elementos do processo KDD auxilia a conceituação de um sistema de extração do conhecimento, uma vez que este sistema seria todo um ambiente integrado que permitiria aos seus usuários realizar o complexo processo de descoberta do conhecimento. Muitos dos sistemas KDD, hoje disponíveis no mercado, estão preparados para atuar em parte do processo KDD, oferecendo implementações para a mineração de dados e avaliação de padrões (Brachman & Anand, 1997),(Frawley *et al.*, 1992).

Dos modelos analisados neste trabalho, independentemente da forma em que cada modelo estabelece o fluxo de trabalho desde a exploração inicial dos dados até a disponibilização do conhecimento para seus usuários finais, três objetivos principais se tornam evidentes:

- seleção e evolução dos dados para um modelo de aprendizagem;

- uma etapa que se destina à utilização de ferramentas para a análise de dados;
- homologação e divulgação dos resultados.

Com base nos objetivos destacados acima, as seções seguintes apresentam os elementos metodológicos e conceituais do processo KDD.

2.2 - Definição do Processo de Extração de Conhecimento em Base de Dados

Para auxiliar no entendimento do processo de Extração de Conhecimento em Banco de Dados, ou *KDD*, pode-se recorrer a algumas definições. Inicialmente, pode-se resumidamente definir um processo em termos de um conjunto de etapas, que ao serem executadas dentro de uma sequência lógica, permitem que se obtenha um produto ou que se execute um serviço. Desta forma, um processo de extração de conhecimento em banco de dados implica que o KDD será composto de um conjunto de passos que envolvem a preparação dos dados, a busca de padrões e a avaliação do conhecimento, que podem ser executadas em diversas iterações. Frawley, Piatetsky-Shapiro e Matheus (Frawley, Piatetsky-Shapiro & Matheus, 1991) descreveram o processo KDD como sendo “..um processo não trivial de identificação de padrões válidos, novos e potencialmente úteis em dados.”, Williams e Huang (Williams & Huang 1996) elaboraram um modelo conceitual de processo KDD, apresentando como elementos primordiais: os dados; o conhecimento prévio e os padrões descobertos; o(s) método(s) de avaliação dos padrões e as operações necessárias para os diferentes estágios do processo. O modelo básico seria dado pelos quatro elementos $\langle \mathcal{D}, \mathcal{L}, \mathcal{F}, \mathcal{S} \rangle$ que consistiam em: o banco de dados ($\mathcal{D}$), a linguagem de representação do conhecimento ($\mathcal{L}$), a função de validação do padrão ($\mathcal{F}$) e as operações ($\mathcal{S}$).

- **Banco de dados ($\mathcal{D}$):** no modelo conceitual de processo de Williams & Huang (1996), o banco de dados é representado como um componente de dados, podendo assumir diversas formas de acordo com o estágio do processo: base de dados de produção, *source data* (derivado do primeiro processo de extração) ou *working data* (dados após o tratamento de pré-processamento e que são utilizados no processo de mineração de dados).

No contexto particular do *source data*, é comum ter um entendimento do banco de dados como uma *relação universal* que mapeia de forma efetiva um conjunto de tabelas em um único repositório, por meio dos relacionamentos entre as mesmas (Ullman, 1988). A relação universal de Ullman tem como objetivo fornecer uma interface unificada para uma variedade de operações que são executadas ao longo do processo de extração de conhecimento.

- **Linguagem de representação do conhecimento ($\mathcal{L}$):** a utilização de uma linguagem para representação do conhecimento, tanto descoberto como prévio, tem como objetivo facilitar a sua interpretação tanto por especialistas, como por sistemas baseados em conhecimento. O modelo conceitual de Williams e Huang adota um conjunto de regras como padrão de linguagem. Entretanto, deve-se levar em conta que alguns cenários mais complexos necessitam de outros padrões de linguagem.

- **Funções de avaliação de padrões ($\mathcal{F}$):** são utilizadas para avaliar o grau de interesse do conhecimento extraído sob o ponto de vista do usuário final. A avaliação dos padrões são importantes na redução do espaço de busca (Holsheimer et al, 1995). As funções de validação de padrões mapeiam expressões ou regras de produção dentro de um conjunto de valores numéricos. Um padrão também pode ser avaliado pela sua utilidade, inovação e validade dentro do domínio de aplicação.

- **Conjunto de operações (S):** uma variedade de operações pode ser utilizada durante todo o processo de extração do conhecimento. A escolha de uma operação depende da maneira em que um problema foi formulado e como as soluções são interpretadas pelo usuário. Williams e Huang (Williams & Huang, 1996) identificaram as seguintes operações que podem ser empregadas durante o processo KDD:

1. Operações para extração de Amostras em banco de dados;
2. Operações de Pré-Processamento de Dados;
3. Operações para Meta-Data;
4. Operações de Mineração de Dados;
5. Operações de Predição;
6. Operações de Visualização.

2.3 - Modelos de Processo KDD

Dentre os modelos de processo de extração de conhecimento investigados neste trabalho, algumas variações podem ser destacadas quanto à composição das fases ou etapas necessárias para concretizar o processo com a disponibilização da informação para o usuário final. As diferenças entre os modelos se evidenciam principalmente quanto à separação das etapas a partir de atividades consideradas mais importantes, ou então, com agrupamentos de atividades em fases, cujo marco é estabelecido pela mineração de dados. Dentre os modelos analisados, todos revelaram flexibilidade quanto à divisão do projeto de extração em iterações, e com grande interatividade com o usuário final.

Fayad, Piatetsky-Shapiro e Smyth (Fayyad, Piatetsky-Shapiro & Smyth, 1996) elegem atividades consideradas essenciais, revelando maior quantidade de etapas que não necessitam de particionamentos. Neste trabalho, optou-se pela separação das etapas do processo de extração de conhecimento

agrupando-as conforme seus objetivos em relação à mineração de dados, estabelecendo assim, três etapas principais: o pré-processamento, a mineração de dados e o pós-processamento, conforme descrito na tab. 2.1. Essa forma de agrupamento de atividades foi proposta por Baranauskas (Baranauskas, 2001).

Tab. 2.1: Distribuição de atividades do processo KDD por fases.

<i>Fases</i>	<i>Etapas da fase</i>
<i>Pré-processamento</i>	Compreensão da Área de Domínio do problema
	Compreensão do Modelo de Dados
	Seleção de Dados
<i>Mineração de Dados</i>	Seleção de Método de Mineração de Dados
	Planejamento da Geração de Modelos e Testes de Hipóteses
	Aplicação de Técnicas de Mineração de Dados
<i>Pós-processamento</i>	Visualização e Interpretação de Padrões Minerados e avaliação.

2.3.1 - Modelo de Fayad, Piatetsky-Shapiro e Smyth

O processo KDD é caracterizado pelas diversas interações com os usuários, sendo realizadas em vários pontos do processo de extração, e ainda, com etapas organizadas por iterações. Segundo um modelo idealizado por Fayad, Piatetsky-Shapiro e Smyth (Fayad, Piatetsky-Shapiro & Smyth, 1996), o processo KDD pode ser definido por nove etapas, abaixo destacadas:

1. Compreensão do domínio: envolve entender de forma ampla a área de domínio, priorizando os objetivos do usuário final.
2. Seleção de dados: criar um conjunto de amostras que serão alvo do processo de extração. Durante esta atividade, executa-se uma definição das sentenças de acesso ao banco de dados para a extração dos registros que satisfazem à condição de pesquisa.
3. Limpeza e pré-processamento: envolve executar as operações básicas para a remoção de dados irrelevantes para o processo de extração, mantendo apenas os conjuntos de registros necessários. A eliminação de dados *missing* também faz parte dessa etapa.
4. Redução e projeção: encontrar características para representar a dependência entre dados é o principal objetivo dessa tarefa. Com a redução da dimensionalidade e métodos de transformação, o número efetivo de variáveis em consideração é geralmente reduzido.
5. Seleção da tarefa de mineração de dados: escolher a tarefa de mineração de dados que será aplicada: classificação, regressão, *clustering*, etc.
6. Escolha do algoritmo de mineração de dados: escolher métodos para verificação de padrões em dados.
7. Mineração de dados: identificar padrões em uma representação particular ou um conjunto de representações: regras por classificação ou árvores de decisão, regressão, *clustering* e assim por diante.
8. Intepretação dos padrões minerados: utilizar uma ferramenta específica para a visualização.
9. Incorporação ou consolidação: nessa etapa, ocorre a incorporação do conhecimento em recursos para o usuário final, podendo ser por agregação em bases de conhecimento ou sistemas que utilizam esses padrões em bases de regras, ou simplesmente, disponibilizando esse conhecimento em ferramentas específicas ou relatórios.

2.3.2 - Modelo de Williams e Huang

Ainda no ano de 1996, um trabalho divulgado por Williams e Huang (Williams & Huang, 1996) resumia o processo KDD, simplificando o processo de Fayad, Piatetsky-Shapiro e Smyth. Nessa nova proposta, foram definidas quatro etapas, com uma separação bem definida entre os elementos de entrada para cada etapa do processo e o produto de cada etapa, conforme descrito abaixo:

1. Seleção de Dados,
2. Pré-processamento,
3. Mineração de dados,
4. Avaliação,

Essa definição permite uma compreensão satisfatória do processo para os usuários finais, auxiliando no processo de gestão do processo ao deixar claro os pontos limites que separam as etapas, bem como o produto final de cada etapa. A tabela 2.2 detalha as etapas propostas nesse modelo.

Tab. 2.2: Processo KDD segundo Williams & Huang.

<i>Etapas</i>	<i>Descrição</i>	<i>Elemento de Entrada</i>
<i>Seleção de Dados</i>	Extração de conjunto de dados representativos do problema definido pelo usuário	Dados brutos do banco de dados
<i>Pré-Processamento</i>	Limpeza e formatação dos dados no padrão necessário para a ferramenta de mineração de dados.	Conjunto de Dados representativos do problema
<i>Mineração de Dados</i>	Aplicação de ferramentas de mineração para identificação de padrões	Dados de trabalho
<i>Avaliação</i>	Avaliação dos padrões extraídos	Padrões

2.2.3 - Modelo de Hipp, Güntzer e Nakhaeizadeh

Uma recente investigação avaliou a aderência da mineração de regras de associação a um processo de extração de conhecimento iterativo e iterativo para cenários complexos (Hipp, Güntzer & Nakhaeizadeh, 2002). Aquela pesquisa disponibilizou um modelo de processo elaborado a partir de trabalhos de Chapman *et al.* (Chapman *et al.*, 2000) e Wirth e Hipp (Wirth & Hipp, 2000). Na proposta de Hipp, Güntzer e Nakhaeizadeh, o modelo de processo de extração de conhecimento consiste nas seguintes etapas:

- Compreensão do domínio: na primeira etapa do processo KDD se estabelece uma visão geral da área de negócio. O objetivo dessa fase é estabelecer com o usuário os objetivos do projeto de extração de conhecimento, de acordo com os problemas apresentados pelo usuário. Nesta etapa, há um estudo dos processos de negócio que geram as informações ao usuário.
- Compreensão dos dados: nesta etapa, os analistas de mineração de dados buscam compreender o modelo de dados a partir do processo de negócio elaborado juntamente com o usuário, identificando os repositórios de dados e os atributos relevantes para o processo de preparação de dados.
- Preparação dos dados: nessa etapa ocorre a composição do conjunto de dados que serão submetidos ao algoritmo de mineração de dados. Esta etapa cobre os aspectos sintáticos, como a formatação e transformação de dados para a mineração, e os aspectos semânticos, como a seleção de tabelas, registros e atributos do banco de dados.
- Modelagem: nesta etapa ocorre a mineração de dados, aplicando-se os algoritmos de mineração de dados conforme a avaliação do problema de negócio, previamente identificado, e o conjunto de dados selecionados e tratados nas etapas anteriores.
- Avaliação: a avaliação dos resultados obtidos a partir da mineração de dados pode ocorrer em três aspectos: o primeiro aspecto seria assegurar que a

aplicação da técnica de mineração de dados foi bem sucedida. Por exemplo, pode-se questionar se o algoritmo de mineração de dados foi capaz de ler e interpretar corretamente os dados preparados. Ou ainda, se toda a informação necessária foi fornecida ao algoritmo. O segundo aspecto seria quanto à necessidade de se investigar se os resultados obtidos são realmente robustos. Alguns métodos dão suporte direto a esta decisão, ao calcular algumas medidas de significância, enquanto que outros deixam este aspecto totalmente sob a responsabilidade do analista. O terceiro aspecto desta fase de análise é averiguar se todos os pontos relevantes ao negócio foram de fato considerados.

- Deployment: nesta etapa ocorre a transferência dos resultados obtidos ou a disponibilização dos padrões extraídos ao ambiente de negócio.

2.4 - Organização do Processo de Extração de Conhecimento

Embora seja crescente o interesse das corporações na utilização de processos de extração de conhecimento a partir das bases de dados, nem sempre o conhecimento extraído se apresenta útil para o usuário final. Isso ocorre porque nem sempre o conhecimento extraído atende às necessidades do usuário, ou então, o conhecimento extraído pode ser redundante ou inconsistente com um conhecimento *a priori* que o usuário possui. Processos de extração de conhecimento dependem de fatores críticos, tais como as características do modelo de dados, a garantia de consistência dos dados e o quanto o conhecimento acerca do negócio e sua evolução está esclarecida como uma necessidade essencial para o usuário final (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Desta forma, surge a necessidade de se organizar as etapas e atividades do processo de extração de conhecimento, estabelecendo um acordo entre especialistas das áreas de negócio e especialistas em mineração de dados, que devem colaborar em várias etapas ao longo do processo. Um

processo de extração de conhecimento define um plano de ações, um modelo do que fazer em cada fase, elencando e organizando as atividades essenciais dentro de cada etapa do processo de extração do conhecimento. Deve-se lembrar também que muitas das dificuldades na implementação de um processo de extração de conhecimento reside na ausência de avaliação da base de dados e como ela está distribuída fisicamente em termos de ambiente e servidores. É comum, até mesmo em empresas de médio porte, encontrar soluções baseadas em softwares para cada parte de um processo organizacional. Isto significa que uma informação pode ser tratada por algumas aplicações, cada qual com sua base de dados, gerando informações de forma distribuída e nem sempre dentro de um padrão comum. Para esses casos, integrar esses dados dentro de um padrão que atenda às necessidades da extração de conhecimento não é uma tarefa trivial. Isso exige um acompanhamento dos especialistas do domínio para um mapeamento dos processos que geram essas informações na etapa de pré-processamento.

Os requisitos para a aplicação de um processo KDD podem variar de acordo com o domínio da aplicação, e mesmo dentro de um mesmo domínio podem ocorrer algumas variações de acordo com o problema abordado. Em aplicações de apólices de seguros, por exemplo, a extração de regras para o ajuste do valor do prêmio e a detecção de fraudes necessitam de processos KDD totalmente diferentes, mesmo que o conjunto de dados sejam semelhantes ou equivalentes. No cenário atual, com uma grande diversidade nos modelos de bancos de dados e diferentes necessidades dentro de um mesmo domínio do conhecimento, a tecnologia KDD passou a ser considerada como um processo empírico de tentativa-e-erro. Por essa razão, este trabalho apresenta uma orientação formal para organização do referido processo, de tal forma que se tenha uma metodologia eficaz como suporte aos participantes do processo de extração.

Tecnicamente, um sistema de extração de conhecimento em banco de dados é um ambiente integrado, que de alguma forma permite ao usuário executar o complexo processo da extração de conhecimento. As informações

de saída obtidas em um processo de extração de conhecimento têm uma abrangência maior quando comparadas às saídas oferecidas por um sistema de extração, que muitas vezes são resultados de um conjunto restrito de dados. Isto ocorre porque o resultado de um processo de extração de conhecimento permite, por exemplo, especificar na totalidade uma aplicação que atenderá a uma área de negócio. Tal aplicação pode ser desenvolvida e instalada em um ambiente de negócio, oferecendo funções de análise e suporte à tomada de decisão, com base em dados de entrada. O usuário desse tipo de sistema pode eventualmente ser um profissional especialista da área de negócio, cuja interação com o sistema se dá de maneira diferente, pois necessita das ferramentas de extração de conhecimento como auxílio na busca e identificação de eventos relevantes nos dados de negócio, a identificação de tendências e padrões. Neste mesmo contexto, um sistema que monitora um conjunto de registros de uma base de dados e dela extrai padrões pode ser melhor definido como uma aplicação de suporte ao processo de extração de conhecimento, e não um sistema de extração de conhecimento.

2.5 - Ferramentas de suporte ao processo KDD

A implementação de um processo KDD depende de uma série de fatores internos e externos à uma organização para que se obtenha o sucesso pretendido. Por exemplo, implementar a captura de informações relevantes para o processo de decisão nos sistemas transacionais ou a revisão constante dos processos de armazenamento prevendo o controle de qualidade dos dados e minimizar as perdas de dados ocorridas nos processos operacionais. Um fator externo para a implementação dos processos KDD é a colaboração entre as empresas e instituições no fornecimento de dados, por exemplo, empresas que fornecem equipamentos de rastreamento de veículos podem atuar em conjunto com empresas que fornecem seguros de transporte de cargas para identificar um padrão nos furtos de carga. Alcançar a produtividade na identificação de padrões não é uma tarefa que pode ser concluída sem que haja uma interação

entre elementos humanos e técnicos. Muitas organizações levam até alguns anos para obter um ambiente propício para iniciar seus projetos de mineração de dados (Hipp, Güntzer & Nakhaeizadeh, 2002). Na figura 2.3, apresenta-se um quadro resumido da evolução dos recursos tecnológicos que oferecem suporte aos processos decisórios e são fatores essenciais nos projetos de mineração de dados.

Fig. 2.3: Evolução dos recursos de tecnologia da informação

Um dos elementos que se configuram como essenciais para a implementação de um processo KDD é o *data warehouse* (Tan, Steinbach & Kumar, 2005). Embora não esteja implícito nos modelos KDD analisados neste trabalho, a organização dos dados nesse padrão interfere diretamente nos custos dos projetos de mineração de dados, uma vez que essa abordagem de organização dos registros leva em consideração técnicas de armazenamento, manutenção, e recuperação de dados históricos, de forma a tornar seu acesso fácil e rápido. Além disso, as técnicas de mineração de dados parte do princípio da manipulação de um grande volume de dados e sendo assim, um projeto que prevê a extração de dados a partir de um sistema transacional pode causar impactos não desejados em função do acesso concorrente com sistemas operacionais em produção.

Os *data warehouse*, por outro lado, integram múltiplas fontes de dados, gerenciando problemas associados aos diferentes formatos dos dados, suporte a vários sistemas de gerenciamento de bancos de dados, fontes de dados distribuídos, e unificação da representação dos dados e limpeza (Fayyad & Stolorz, 1997). Ao contrário dos bancos de dados que são modelados para apoiar sistemas de suporte às atividades organizacionais cujos dados são atualizados de forma rápida e constante, com volume restrito e controlado, os *data warehouse* são repositórios de dados históricos, com alimentação feita em lotes, alcançando muitas vezes um volume de vários *gigabytes* ou até *terabytes*. Os *data warehouse* tornaram possível o acesso a dados históricos de forma rápida e segura, tornando-se, assim, uma ferramenta muito útil nas atividades de suporte à decisão. A aquisição dos registros históricos que alimentam os *data warehouse* muitas vezes não se limita a uma fonte única, necessitando de implementações de rotinas que além de efetuarem a extração, atuam na padronização de registros, eliminando os dados inválidos e inconsistentes. Além disso, o padrão de modelagem de um *data warehouse* é a modelo dimensional, com dados não normalizados, ou seja, constituído de poucas tabelas com muitas colunas, tornando rápido o acesso a um grande volume de dados. O modelo dimensional mais utilizado é o *star schema*, com uma tabela central, conhecida como tabela de fatos, e demais tabelas de suporte, conhecidas como tabelas de dimensão.

2.6 - Pré-Processamento de Dados

A etapa de pré-processamento consiste na aplicação de procedimentos sequenciais na extração e tratamento de dados até que estes estejam em um formato adequado para a mineração de dados. Esta etapa também é conhecida como Preparação de Dados, e por envolver uma série de atividades até a sua finalização, que envolve inclusive o estudo de processos, acaba se tornando a etapa que exige maior esforço dentro de um projeto de extração de conhecimento.

Os dados de entrada utilizados nessa fase precisam ser identificados e avaliados de acordo com as peculiaridades de cada caso. Em muitas ocasiões, o analista que irá atuar no processo de extração de conhecimento não possui familiaridade com a qualificação dos dados a serem extraídos, e por isso, aqui são propostas atividades iniciais de levantamento do domínio do conhecimento e avaliação do modelo de dados que recebe as informações processadas.

A partir da atividade referente ao estudo do modelo de dados é que serão planejadas as extrações de dados brutos e seu tratamento, o que inclui a seleção de dados, as operações de limpeza de dados para correção de valores vagos e a eliminação de redundâncias. Além das atividades de tratamento de dados, é necessário, muitas vezes, aplicar procedimentos de correlação de valores, pois o armazenamento de dados em repositórios distribuídos é muito comum, que ocorre quando várias bases de dados precisam ser consolidadas e um mesmo conceito pode assumir tipos de valores diferentes.

Os dados de entrada (*input*) podem estar em vários formatos (arquivos de log's, planilhas, tabelas em banco de dados relacional ou dimensional), centralizados em um único repositório ou então, distribuídos em vários recursos de armazenamento, inclusive em bancos de dados de sistemas transacionais. A proposta da etapa de *pré-processamento* é fazer com que os dados de entrada sejam convertidos em um formato apropriado e uniformizado para uma análise posterior. Nesta etapa, os dados brutos são selecionados, servindo assim como referência para a mineração de dados. Após uma seleção inicial dos registros relevantes ao processo, os registros resultantes são novamente selecionados, agora somente analisando aspectos de completeza dos dados, excluindo aqueles dados incompletos, duplicados, imprecisos e exeções. A seleção dos dados deve ser acompanhada por um especialista do domínio, pois este detém o conhecimento sobre o processo ou uma cadeia de processos operacionais que geram a informação e apóiam as atividades organizacionais, porém sua participação não deve ser estendida na qualificação dos atributos da análise de dados, pois pode comprometer a análise, rementendo a pontos não relevantes para o usuário final (Berry & Linoff, 2004).

O pré-processamento deve eliminar a diferença de tipos nas variáveis que representam um mesmo conceito, ou seja uniformizar os atributos, que muitas vezes foram extraídos de bancos de dados distintos. Por exemplo, um atributo pode ser definido com caracteres de texto num determinado banco de dados, enquanto que esse mesmo atributo, proveniente de outro banco, pode assumir valores numéricos. Com a finalização desta etapa, os dados são transformados, de modo a facilitar a utilização das técnicas de *mineração de dados*.

Este trabalho propõe o estudo dos processos organizacionais e as atividades que as amparam. Muitas empresas, de diversos ramos de atuação, adotam o mapeamento de seus processos mais importantes por meio de representações gráficas, utilizando uma notação padrão. Dentre as notações mais utilizadas na representação dos processos, as notações IDEF0 e UML (com o Diagrama de Atividades) fornecem uma compreensão satisfatória do processo para o analista que irá atuar na extração do conhecimento e com pouco conhecimento sobre a área de negócio.

Quando a área não dispor desse tipo de documentação, ou quando a informação não estiver disponível, iniciar um trabalho de representação dos processos utilizando uma dessas notações não é uma tarefa difícil de ser implementada, principalmente pelas vantagens que agrega ao projeto:

1. Facilita a comunicação entre os participantes por oferecer uma notação simples e padronizada;
2. Oferece uma boa visão das atividades mais importantes, as entradas/saídas e recursos de apoio.
3. Propicia considerável disponibilidade de ferramentas que suportam essas notações e que podem ser adquiridas sem custo.
4. A integração com outros processos é facilmente percebida e agregada à visão do processo principal.

A notação IDEF0 fornece um conjunto de elementos gráficos que representam as atividades do processo, as informações de entrada e saída, as regras ou restrições das atividades e as ferramentas de apoio, conforme exemplo da fig. 2.4. O detalhamento das atividades é feita em níveis inferiores – vertical ao processo macro ou principal.

Fig. 2.4: Modelo de Processos, com atividades, fluxos, entradas e saídas.

Por outro lado, a notação UML (*Unified Modeling Language*), muito utilizada na modelagem visual de softwares orientados a objetos, fornece elementos para a representação de processos com o Diagrama de Atividades. O detalhamento do Diagrama de Atividades ocorre pela separação das responsabilidades de um grupo de atividades por meio das “raias”, conforme descrito na figura 2.5. O nível de detalhamento do Diagrama de Atividades é um pouco maior em relação ao mapa de processos do IDEF0, por tratar também das exceções, restrições e desvios dentro de um mesmo nível, além de possibilitar a representação de atividades que ocorrem de modo concorrente.

Fig 2.5: Diagrama de Atividades: modelo de processo com UML.

Após o entendimento do processo organizacional que será estudado, o analista deverá planejar o projeto de extração de conhecimento, distribuindo as atividades dentro de etapas conforme o modelo KDD que será adotado. Nessa

etapa, assume-se como pré-requisito um conhecimento geral sobre os problemas encontrados pela área e as dificuldades causadas pela ausência ou insuficiência de recursos de suporte à decisão.

O conhecimento sobre o processo organizacional é o ponto de partida para o projeto KDD, mas não é suficiente para iniciar as atividades de manipulação de dados da etapa de pré-processamento. A atividade inicial de elucidação do processo organizacional deve ser complementada por um estudo da base de dados disponível, seja ele de um sistema transacional ou repositório histórico centralizado – o data warehouse. Uma forma de se buscar um entendimento das bases de dados é a partir dos modelos de dados. Os modelos de dados são criados na fase inicial dos projetos de sistemas transacionais ou durante a implementação de um data warehouse, sendo que podem representar a visão lógica ou física de um sistema (Kimball, 1996). Na visão lógica, cada unidade de dados representa uma entidade, conceito ou objeto e mantém informações das entidades também conhecidas como atributos. Por exemplo, Pedido e Conta Corrente são exemplos de entidades ou objetos e número_pedido, data_pedido, código_agencia e nome_agência são informações ou atributos das entidades. A interação entre entidades dentro de um modelo conceitual é conhecido como relacionamento. A visão física de um modelo de dados é a representação da estrutura de dados tal como foi implementada em um servidor de banco de dados, dessa forma, a representação física de uma entidade num banco de dados é a tabela, a representação de um atributo é a coluna e os relacionamentos são representados por índices ou chaves.

Os modelos de dados dimensionais são modelos de dados que representam a estrutura de uma base de dados históricos, do tipo *data warehouse*. Em termos de notação, o modelo de dados dimensional também exibe as entidades de dados e os relacionamentos entre essas entidades. Um fator relevante está na estrutura dessas entidades, normalmente, uma entidade de um modelo dimensional agrega mais atributos que uma entidade de um modelo transacional, e a entidade que mantém as movimentações, também conhecida como tabela de fatos, relaciona-se com outras entidades a partir dos

atributos de dimensão. Um exemplo de modelo dimensional é exibido na figura 2.6. Nesse modelo, observa-se que a entidade central, a entidade de fatos, relaciona-se com outras entidades a partir dos atributos de dimensão. Esse tipo de modelagem também é conhecido como modelo estrela, sendo o mais aplicado na implementação de *data warehouse* por manter uma estrutura que facilita o acesso e a recuperação de um grande volume de dados (Kimball, 1996).

Fig.2.6: Modelo de dados dimensional tipo estrela.

2.7 - Seleção de Dados

A seção 2.3 apresentou os modelos de processos de extração de conhecimento que podem ser aplicados nas organizações em diferentes áreas de atuação. Também foi proposto um trabalho que enfatiza duas atividades prévias no pré-processamento de dados, que permite ao analista de mineração de dados a familiarização com processos de negócio e os termos e conceitos

específicos da área de domínio que será o foco das atividades de mineração, além da avaliação da base de dados que foi modelada para atender tais atividades organizacionais. Essas atividades que iniciam o pré-processamento permitem uma boa avaliação da quantidade de informações relevantes presentes ou não nas bases de dados das organizações.

O sucesso da mineração de dados depende não apenas da escolha do algoritmo de aprendizagem mais adequado e a aplicação desse algoritmo sobre os dados, mas também de critérios de seleção, dentre outros. Os algoritmos necessitam de parâmetros de entrada para o processamento e estes necessitam dos valores apropriados. Em muitos casos, o desempenho do algoritmo de aprendizagem pode ser melhorado de acordo com a escolha apropriada dos valores disponíveis (Witten & Frank, 2005).

Esta seção discute algumas abordagens rotineiramente adotadas para o processo de seleção de dados. Diferentemente das atividades iniciais que envolvem a compreensão do domínio ao qual serão extraídas as informações, nesta etapa ocorrem, de fato, a obtenção dos dados e sua manipulação, a partir de abordagens e técnicas que podem ser aplicadas isoladamente ou encadeadas, conforme as características dos dados extraídos dos bancos de dados. As principais abordagens para a atividade de seleção de dados são (Tan, Steinbach & Kumar, 2005), (Witten & Frank, 2005):

- Agregação,
- Amostragem,
- Redução de dimensionalidade,
- Seleção por atributos ou critério,
- Criação de atributos,
- Discretização e binarização,
- Transformação de variáveis.

Em linhas gerais, pode-se afirmar que essas técnicas são utilizadas com duas finalidades bem distintas: a seleção de objetos de dados e atributos para análise ou a criação e mudança de atributos. Em ambos os casos, o

objetivo final é melhorar o processo de extração de conhecimento em termos de prazo, custo e qualidade. A seguir, cada uma dessas atividades será apresentada detalhadamente:

- *Agregação*: consiste na combinação de dois ou mais objetos de dados em um único objeto, permitindo o agrupamento de um conjunto de registros, com base em algum critério ou atributo. Em termos práticos, a agregação permite uma melhor visualização de indicadores implícitos em valores de atributos. Um exemplo de aplicação da agregação ocorre na representação dos dados referentes às transações de vendas de uma determinada cadeia comercial, que possui estabelecimentos localizados em vários estados, e que armazena em um banco de dados milhares de transações de vendas por dia. Uma forma de agregar essas transações seria representá-las em um único registro utilizando como referência um atributo que represente o total de vendas por estabelecimento. Dessa forma, o total de objetos de dados estaria reduzido ao número de estabelecimentos.

- *Amostragem*: é uma abordagem muito comum para a seleção de um subconjunto de objetos de dados a serem analisados. Basicamente, a seleção de amostras em bases de dados ocorre com a definição mínima de critérios, de forma que se possa obter uma amostra representativa em relação à população. Uma amostra é representativa se os objetos que a compõem compartilham das mesmas propriedades dos objetos de dados que compõem a população de interesse. Essas propriedades são justamente definidas com os critérios, que podem assumir condições booleanas, valores exatos ou faixas de valores.

- *Redução de dimensionalidade*: os registros que compõem uma base de dados podem expressar diversas informações, por exemplo, os dados que representam o preço de fechamento diário de produtos de um estoque num período de 30 anos: nesse caso, o atributo preço para determinados dias, poderá conter milhares de registros. A redução de dimensionalidade é indicada nos processos de seleção para reduzir a quantidade de atributos envolvidos no processo de análise, eliminando informações irrelevantes em relação ao

problema que está sendo estudado. Adicionalmente, a eliminação de atributos irrelevantes traz ganhos no processamento dos algoritmos de mineração, principalmente quando é efetuada a tarefa de classificação de registros. O termo redução de dimensionalidade é reservado para as técnicas para redução de dados, onde novos atributos são criados a partir da combinação de outros atributos.

- *Seleção por atributos ou critério*: uma outra forma de se reduzir a dimensionalidade dos dados é utilizar somente um subconjunto das características dos dados. Essa abordagem é muito aplicada, pois muitas bases de dados mantêm informações redundantes e irrelevantes no contexto do problema. Entende-se que atributos redundantes são aqueles que duplicam a informação contida em um ou mais atributos e os irrelevantes referem-se a aqueles que não são necessários a um projeto de extração de conhecimento em particular.

- *Criação de atributos*: a proposta dessa abordagem é a criação de novo conjunto de atributos que capturem informações importantes de forma efetiva, a partir de atributos originalmente presentes na base de dados. A criação de atributos é bastante comum, pois algumas técnicas de mineração de dados necessitam de atributos em formatos que não estão presentes no banco de dados ou em *data warehouses*.

- *Discretização e Binarização*: alguns algoritmos de mineração de dados, principalmente os classificadores, necessitam que os dados estejam no formato categórico. Outros algoritmos, que buscam por padrões de associação, requerem que os atributos estejam no formato binário.

- *Transformação de variáveis*: refere-se à transformação aplicada a todos os possíveis valores de uma variável e que podem ser de duas categorias: *transformações funcionais simples* e *normalização*. A transformação de dados a partir da normalização padroniza os valores de um atributo dentro de uma escala específica de medida, tal como 0 e 1. A normalização *mínimo-máximo* executa uma transformação linear a partir dos dados originais, onde os valores min_A e

max_A são os valores mínimos e máximos de um atributo A. Dessa forma, para associar um novo valor normalizado v' a um valor conhecido v de um atributo A, utiliza-se:

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{novo_min}_A - \text{novo_min}_A) + \text{novo_min}_A$$

Para aplicar a normalização mínimo-máximo, é importante que sejam conhecidos os valores limites do atributo a ser transformado, o que implica numa primeira extração de dados a partir da base de dados original. Para exemplificar a normalização, podemos adotar o seguinte exemplo: supondo que o atributo `valor_contrato` possua o valor mínimo de \$ 12000, 00 e valor máximo de \$ 98000, a normalização dentro da escala [0,1] para o valor \$ 73600,00 será:

$$v' = [(73600 - 12000) / (98000 - 12000)] \times (1 - 0) + 0 = 0,716.$$

Há várias motivações para a utilização das técnicas de seleção de dados descritas anteriormente. No caso da agregação, conjuntos de dados com menos registros reduzem o tempo de processamento dos algoritmos de mineração de dados. A visualização dos dados em alto nível faz da técnica de agregação uma alternativa viável para a percepção de indicadores de desempenho, e que não são perceptíveis quando o conjunto de dados da análise envolve milhares de transações. Uma desvantagem da agregação é a perda de certos detalhes das transações, pois os registros de movimentação são agrupados em torno de um atributo específico. Ao aplicar a agregação a uma base de dados que contempla as transações de vendas de uma cadeia de lojas, agregar os registros dentro do período de um ano para uma determinada loja, pode “esconder” o desempenho das vendas de um produto em determinados meses ou semanas, já que a agregação, nesse caso, consideraria a média dos valores das vendas de todos os produtos.

Numa abordagem estatística, uma amostra é utilizada tanto na investigação preliminar dos dados quanto na análise final. Embora a mineração de dados também utilize com frequência a seleção de amostras, sua motivação ocorre em contextos diferenciados: no caso dos estatísticos, a utilização da amostra ocorre devido às dificuldades na obtenção da população e custos proibitivos desse processo; enquanto que os analistas de mineração de dados utilizam a amostra por ser demasiadamente custoso e demorado o processamento de toda a população de dados. O princípio chave da utilização da amostragem é garantir a seleção de amostras representativas sobre a população de interesse.

A seleção de atributos pode eliminar quase que imediatamente aqueles irrelevantes ou redundantes dentro do conjunto de atributos analisados, a partir de um certo conhecimento do domínio. Por outro lado, em muitos casos, é conveniente utilizar uma abordagem sistemática para a obtenção do melhor subconjunto de atributos. Uma abordagem ideal para a seleção de atributos seria tentar todas as combinações possíveis de atributos como parâmetros ao algoritmo de teste de relevância e verificar qual subconjunto de atributos apresenta a melhor combinação. A desvantagem desse método é a quantidade de testes necessários, pois para n atributos há 2^n combinações possíveis.

A binarização e a discretização são atividades de transformação de dados cujo objetivo é melhorar o desempenho dos algoritmos classificadores de eventos e os de extração de padrões de associação. Na normalização, o objetivo é obter a padronização das variáveis. Se variáveis de diferentes tipos necessitam ser combinadas de alguma forma, o uso da transformação de variáveis pela normalização evita que variáveis com escala de valores maiores tenham domínio sobre os resultados do cálculo. Como exemplo, pode-se considerar uma comparação entre indivíduos, baseada em suas rendas e faixas etárias. Em termos absolutos, a diferença entre as rendas dos indivíduos será bem maior do que diferença em termos de idade. Se as diferenças entre as escalas de valores de renda e idade não forem examinadas por medidas de similaridade ou dissimilaridade, então, conseqüentemente a variável renda irá se

sobrepor às medidas de idade, dominando o cálculo. A tabela 2.3 destaca os principais elementos considerados na fase de pré-processamento.

Tabela 2.3: Elementos relevantes da fase de pré-processamento.

Elementos do Pré-Processamento		
<i>Etapa</i>	<i>Atividade</i>	<i>Método</i>
<i>Conhecimento do domínio</i>	Análise de processos	Entrevista, estudo do processo – IDF0
<i>Estudo da base de dados</i>	Análise de Modelo de dados e Ambiente	Modelo Entidade-relacionamento
<i>Seleção de Dados</i>	Agregação	
	Amostragem	Amostragem progressiva
	Redução de Dimensionalidade	PCA, SVD
	Seleção por atributos ou critério	Filtro, Wrapper
	Criação de atributos	
	Binarização e Discretização	Discretização Supervisionada e não supervisionada
	Transformação de variáveis	Transformações funcionais simples
		Normalização

Este trabalho destaca a importância da utilização de abordagens que apoiem a fase de pré-processamento de dados, de forma a disponibilizar uma base consistente e adequada para a execução das tarefas de mineração de dados nos padrões necessários para a técnica que deverá ser aplicada no reconhecimento dos padrões e extração de regras.

Observa-se que na organização de um processo de extração de conhecimento, uma atenção especial deve ser dedicada à atividade de seleção de atributos, e na possibilidade de automatizá-la. Para isso, a combinação de técnicas de mineração de dados, aplicadas também nessa fase, podem destacar o melhor subconjunto de atributos, ou ainda permitir o reconhecimento de atributos irrelevantes (Witten & Frank, 2005).

Atualmente, a identificação de atributos irrelevantes no processo de aprendizagem representa um ponto de grande preocupação, principalmente pelo fato de que certos algoritmos de aprendizagem serem mais sensíveis ao trabalharem com registros “locais”. De fato, mostrou-se que a quantidade de registros necessários para produzir um determinado nível de desempenho dos algoritmos de aprendizagem baseados em instâncias aumenta exponencialmente com a quantidade de atributos irrelevantes presentes no contexto dos dados analisados (Frank & Witten, 2005). Já o algoritmo Naive Bayes não fragmenta o seu espaço de amostras, ignorando os atributos irrelevantes, ao assumir que os atributos são independentes entre si. Por outro lado, o Naive Bayes degrada o modelo de decisão se os atributos redundantes forem envolvidos no processo de análise.

Devido ao efeito negativo dos atributos irrelevantes nos algoritmos de aprendizagem, é comum preceder a mineração de dados com atividades de seleção de atributos. Um dos possíveis caminhos para a eliminação desses atributos é a seleção manual, baseada na compreensão do problema e no conhecimento conceitual que se tem dos atributos que compõem a base de extração. Métodos automáticos também podem ser utilizados, de tal forma que a redução da dimensionalidade, suprimindo atributos irrelevantes, melhora o desempenho dos algoritmos de mineração de dados. Esta seria uma atividade mais eficiente em relação à seleção manual, porém necessitaria de uma certa codificação para a seleção do atributo. A redução de dimensionalidade também fornece uma visão mais compacta dos dados, favorecendo a percepção dos atributos mais importantes na visão do usuário.

Na seleção de subconjuntos de atributos, duas abordagens diferentes são essencialmente aplicadas: na primeira, se faz uma avaliação independente dos atributos, baseada nas características dos dados; na segunda, realiza-se o teste de subconjuntos de atributos, utilizando o algoritmo de aprendizagem. A primeira abordagem é chamada de “filtro”, pois o atributo é selecionado para que faça parte do subconjunto de atributos antes de se aplicar o algoritmo de aprendizagem. O segundo método é conhecido como *wrapper*, pois o atributo é incluído no processo de aprendizagem.

A avaliação independente de um subconjunto de atributos é um método interessante no processo de aprendizagem, desde que estejam disponibilizadas técnicas que permitam a verificação de atributos relevantes para a atribuição de uma classe. Um método simples de seleção independente é usar uma quantidade de atributos que permita dividir o espaço amostral de uma forma que se tenha todas as instâncias de treinamento separadas por combinações possíveis de valores dos atributos, o que é possível se o conjunto de treinamento for composto de poucos atributos. Por outro lado, quando a quantidade de atributos aumenta, provavelmente seja mais difícil encontrar registros que se agrupem por terem os mesmos valores em todos os atributos.

As técnicas de mineração de dados podem ser utilizadas para a seleção de atributos, como por exemplo, o algoritmo de árvore de decisão pode ser aplicado a uma base que contenha vários atributos, permitindo somente a seleção dos atributos que são utilizados para a construção da árvore de decisão (na construção da árvore de decisão, atributos não relevantes são abandonados). Esse tipo de seleção pode não ter efeito se o estágio posterior envolver a construção de outra árvore de decisão, mas traz resultados significativos quando combinados com outros algoritmos de aprendizagem. Por exemplo, o algoritmo *nearest-neighbor* é sensível aos atributos irrelevantes e seu desempenho pode ser melhorado com a utilização da árvore de decisão para a seleção de atributos (Frank & Witten, 2005).

Devido à aplicação de uma série de atividades na etapa de pré-processamento, o projeto KDD nesta fase ocupa em torno de 60% até 75% do

tempo total gasto no projeto. A seleção do melhor conjunto de atributos e a transformação dos valores no formato necessário, muitas vezes, exige que essas atividades sejam repetidas e reavaliadas, e dependendo do caso, retorna-se ao programa de extração de dados para a obtenção de atributos não envolvidos na análise anterior. Na etapa de pré-processamento avalia-se todo o cenário de restrições que tornariam a etapa de mineração de dados pouco elucidativa em termos de extração de conhecimento.

3. Mineração de Dados

Introdução

No capítulo anterior, um cenário ainda bastante comum nas corporações foi descrito: a transformação de dados em conhecimento útil usando análises e interpretações pessoais. Embora essas organizações geralmente mantenham ferramentas computacionais ou procedimentos para o acesso a dados históricos, muitas dessas empresas ainda recorrem à análise de dados a partir de planilhas e relatórios. O elemento dificultador nessa abordagem reside na dependência de analistas humanos, que precisam cruzar informações e identificar tendências em um conjunto de informações cada vez maior.

A limitação dessa forma de extração de conhecimento levou ao desenvolvimento de técnicas de extração automática de conhecimento sobre uma grande massa de dados, que aliando técnicas estatísticas de reconhecimento de padrões às teorias da inteligência artificial e sistemas de gerenciamento de banco de dados, culminaram com a nova área da mineração de dados (Tan, Steinbach & Kumar, 2005),(Hosking, Pednault & Sudan, 1997) .

Em linhas gerais, a utilização das técnicas de mineração de dados permite obter uma sumarização eficiente a partir de uma grande quantidade de registros, identificando estruturas e relacionamentos de interesse dentro de um conjunto de dados. Além dessas funcionalidades, as técnicas de Mineração de Dados aplicadas a um conjunto de dados previamente observados permitem a construção de preditores para observações futuras (Hosking, Pednault & Sudan, 1997). Este capítulo aborda os conceitos que envolvem a mineração de dados, as principais técnicas e tarefas da mineração de dados e sua aderência nos modelos de processo KDD.

3.1 - Conceitos de Mineração de Dados

A mineração de dados pode ser definida como um conjunto de procedimentos de extração automática de conhecimento a partir de um grande volume de dados, unindo métodos de análise de dados tradicionais a sofisticados algoritmos para processamento de dados em grande escala (Tan, Steinbach & Kumar, 2005). Na definição dada por Watanabe , o reconhecimento de padrões em dados tem como objetivo identificar estruturas de dados onde se verifica algum tipo de relacionamento com o uso de ferramentas estatísticas e com isso, classificar, agrupar ou associar um registro a um grupo de dados que mantém características em comum (Watanabe, 1985),(Jain et al, 2000).

Durante a execução das atividades de pré-processamento, a disponibilidade dos dados em confronto com o problema apresentado permite qualificar o aspecto da mineração de dados, onde se estabelece a utilização de tarefas direcionadas ou não direcionadas. Na mineração de dados direcionada, o principal objetivo é descrever um conjunto de dados para categorizar seus

elementos; já a mineração de dados não direcionada busca padrão ou similaridades entre os elementos de um grupo, sem o uso de classes ou conceitos pré-estabelecidos (Berry & Linoff, 2004).

Quando nos referimos à mineração de dados e suas definições, o termo *tarefa* costuma ser empregado com certa frequência, porém nem sempre sua definição é exposta claramente. As técnicas de mineração de dados são aplicadas a um conjunto de registros para satisfazer uma abordagem de análise de dados, seja ela classificar, associar, agrupar ou discriminar os elementos de uma amostra e sendo assim, uma tarefa de mineração de dados consiste no objetivo da análise dos dados por meio de uma ou mais técnicas de reconhecimento de padrões (Han & Kamber, 2006). Apresentamos a seguir, os quatro principais objetivos ou tarefas de mineração de dados:

3.1.1- Caracterização e discriminação de classes e conceitos

Os dados contidos em uma amostra podem ser associados a uma classe ou conceito. Por exemplo, quando os registros de transações de compras são utilizados para qualificar um cliente em uma determinada categoria, com base na frequência em que realizam os pedidos ou no valor médio de suas últimas compras, está se estabelecendo um conceito de cliente ou definindo os valores para caracterizar esse cliente. Quando uma massa de dados é analisada, os dados utilizados na mineração de dados devem ser úteis para a descrição de classes e conceitos de forma resumida, precisa e concisa (Han & Kamber, 2006).

Uma das aplicações da mineração de dados envolve a descrição de classes e conceitos a partir da caracterização e da discriminação de dados. Técnicas de mineração de dados aplicadas à caracterização de dados levam a uma sumarização dos atributos mais relevantes que determinam uma classe. Por exemplo, a análise de uma carteira de clientes que obtiveram liberação de crédito: casados, entre 33 e 45 anos, com nível superior.

A mineração de dados para a discriminação de dados tem como objetivo comparar agrupamentos diferentes que se formam em uma amostra de

dados. A discriminação de dados tem como objetivo identificar perfis diferentes de uma classe de dados. Tomando como exemplo dois grupos distintos de clientes de uma loja de produtos de informática: a partir dos tipos de produto consumidos ou pelo valor total gasto, a aplicação de técnicas de mineração de dados pode identificar dois grupos, os clientes que compram computadores, têm idade entre 20 e 45 anos, possuem nível superior ou acima e clientes cuja frequência de compra é superior a um ano, composto por jovens e idosos e que não possuem curso universitário. Na estatística clássica, a análise de discriminação leva um modelo de classificação, quando se tem como objetivo prever o grupo ao qual um indivíduo pertence com base em um conjunto de características ou variáveis mantidas pelo indivíduo (Fisher, 1936),(Tan, Steinbach & Kumar, 2005) .

3.1.2 - Padrões por frequência, associações e correlações entre dados

Os padrões por frequência que são identificados nas bases de dados são ocorrências de dados relacionados em uma base de dados, e podem se apresentar como um *conjunto de itens e padrões sequenciais de dados*. Um padrão por conjunto de itens ocorre, por exemplo, quando se observam ocorrências de valores relacionados dentro de transações em bancos de dados, ou seja, a frequência de pares de dados nos registros de dados. Um exemplo prático muito citado para o entendimento desse conceito é a análise da *cesta de compras* onde, com base em transações de compras, foi possível identificar a frequência de produtos que são comprados aos pares (Berry & Linoff, 2004). O padrão de frequência sequencial leva em consideração a verificação de um relacionamento entre dados pela ordem ou sequência em que aparecem nas transações de dados. O diferencial entre o conjunto de itens e o padrão sequencial é que no primeiro caso, o relacionamento entre dados ocorre em uma transação conceitualmente fechada, por exemplo, um pedido de compra; no caso do padrão sequencial, a contagem do relacionamento ocorre em transações diferentes. Desta forma, pode-se dizer que um padrão sequencial ocorre quando se analisa um conjunto de pedidos de compra para identificar um

comportamento na obtenção de certos itens. Por exemplo, no caso de cestas de produtos em sites de compra online, computadores são geralmente adquiridos antes de uma câmera digital, e assim por diante.

3.1.3 - Análise de agrupamentos de dados ou clustering

A análise de agrupamentos pode ser definida a partir de uma comparação com outra técnica de mineração de dados já conhecida, a classificação. Diferente da classificação, na análise de agrupamentos ou *clustering*, não se conhece previamente as classes que permitem identificar ou selecionar os objetos de dados. Dessa forma, os agrupamentos são utilizados justamente para auxiliar no processo de identificação das classes, pois os objetos de dados são agrupados baseados no princípio da *máxima similaridade intra-classe* e *mínima similaridade inter-classe*, ou seja, os agrupamentos são compostos por objetos que mantêm alta similaridade em relação a outro objeto do mesmo grupo, porém baixa similaridade com objetos pertencentes a grupos diferentes (Jain & Dubes, 1988).

Nos processos de extração de conhecimento em banco de dados, a análise de agrupamentos pode ser utilizada com outra tarefa de mineração de dados em função da necessidade de identificação de classes ou revisão das classes já existentes. A análise de agrupamentos pode identificar as classes que posteriormente serão utilizadas em modelos de classificação.

3.1.4 - Classificação e predição de eventos

A classificação é um processo de identificação de um modelo que permite descrever e distinguir classes de dados ou conceitos e que posteriormente podem ser utilizados para predições de classes. O processo de criação dos modelos de classificação utilizam um conjunto de dados conhecidos como base de treinamento. Após a identificação do modelo, seu desempenho nas classificações é avaliado utilizando-se um conjunto de dados conhecido como base de teste.

O produto da mineração de dados é conhecido como modelo de aprendizagem. Esses modelos são obtidos a partir da aplicação de técnicas de mineração de dados conforme uma tarefa ou objetivo, como descrito anteriormente. A análise de dados realizada por meio de tarefas direcionadas e não direcionadas permitem a geração de modelos que descrevem o comportamento de um conjunto de dados. Esses modelos são separados em dois grupos: os modelos preditivos e os modelos descritivos (Tan, Steinbach & Kumar, 2005):

- **Modelo Preditivo:** tem como objetivo categorizar um registro, ou seja, identificar o valor de um atributo particular com base nos valores de outros atributos. O atributo a ser previsto é também conhecido como variável dependente, enquanto que os atributos utilizados na predição são conhecidos como variáveis independentes (Tan, Steinbach & Kumar, 2005).
- **Modelo Descritivo:** têm como objetivo deduzir padrões - correlações, tendências, agrupamentos, trajetórias e anomalias - que sumariem relacionamentos ocultos nas bases de dados. As tarefas descritivas são exploratórias e frequentemente requerem técnicas de pós-processamento para validar e esclarecer os resultados.

A figura 3.1 destaca um micro-processo que ocorre na etapa de mineração de dados, onde a definição do problema e a massa de dados disponíveis para análise definem a tarefa e a técnica de mineração de dados. Uma vez que o modelo de aprendizagem é gerado, o mesmo será submetido a procedimentos de avaliação de desempenho. Neste trabalho, direciona-se os estudos da aplicabilidade das técnicas de classificação. Desta forma, os tópicos apresentados a seguir enfocam essa abordagem.

Fig. 3.1 – Elementos básicos da mineração de dados.

Em termos práticos, a fase de mineração de dados leva a compor um micro-processo, que pode ser aplicado em várias iterações conforme a necessidade e complexidade do problema, tal como proposto pelo modelo CRISP-Hipp (Hipp, Güntzer & Nakhaeizadeh, 2002).

3.2 - Modelagem Preditiva: Técnicas de Classificação

No dia-a-dia de diversos setores corporativos, a tomada de decisão sobre um grupo de informações novas é uma atividade crítica e estratégica, e ao mesmo tempo deve ser executada de forma eficiente e rotineira. Como exemplo dessas atividades, pode-se mencionar o suporte dos métodos de classificação usados por especialistas de negócios na identificação de tendências no mercado financeiro para a comercialização de ativos (Apte & Hong, 1996); os procedimentos de identificação automática de objetos celestes em bancos de dados de imagens (Fayyad, Djorgovski & Weir, 1996); a análise de empréstimos concedidos por financeiras para classificação de novas concessões; análise de atendimentos de emergência para a organização de turnos de atendimento e

escalas de plantonistas em hospitais; dentre outros. Diante desses elementos, é desejável que os sistemas de suporte à decisão incorporem entre suas funcionalidades a capacidade de classificar eventos, tornando possível, ao usuário final, a categorização de um grupo de informações.

Na Mineração de Dados, a tarefa utilizada para associar um registro a uma categoria é a Classificação, que é basicamente efetuada quando se submete um vetor de atributos a um algoritmo classificador, conforme mostrado na fig. 3.2. O objetivo da classificação é eleger a classe mais provável à qual um registro pertence. Muitos sistemas que implementam rotinas de classificação, dependem do suporte de profissionais com grande experiência para incorporar regras na programação de rotinas de categorização de objetos, nesses sistemas as categorias de objetos devem ser previamente identificadas, muitas vezes de forma manual e eventualmente sujeitos ao problema da base “nula” para um registro que representa uma categoria não prevista .

No processo KDD, a classificação de registros é efetuada na fase de mineração de dados, e permite gerar um modelo descritivo dos registros submetidos ao algoritmo de classificação. Uma vez que o modelo descritivo é validado e passa a ser utilizado em uma base de conhecimento, o modelo preditivo pode ser aplicado para a classificação de novos registros.

Fig. 3.2: Esquema básico de um classificador.

A tarefa de classificação inicia-se com um conjunto de registros, ou amostra, também conhecidos como dados de trabalho, treinamento ou *input*. Os dados de treinamento são obtidos na etapa de pré-processamento (fig. 3.3) e disponibilizados em um banco de dados específico para a mineração dos dados, já em formatação apropriada. Dentre as técnicas utilizadas na classificação, pode-se destacar as Árvores de Decisão, a Classificação baseada em Regras, Redes Neurais, a Classificação por Vizinho Mais Próximo e uma abordagem utilizando o Teorema de Bayes – mais conhecido como Naïve Bayes. Todas as técnicas acima descritas implementam um **algoritmo de aprendizagem** para identificar um modelo que melhor representa o relacionamento entre o conjunto de atributos e uma determinada classe (Hand, 1981),(Weiss & Kulikowski, 1991),(McLachlan, 1992).

Fig. 3.3: Esquema básico de obtenção do modelo de classificação e predição de classes.

3.4 - Modelos de Classificação Baseados em Regras

A técnica de mineração de regras de classificação é amplamente utilizada como método de extração de conhecimento para geração de modelos de aprendizagem. Para a extração das regras de associação, consideramos um ambiente de dados D , que contém registros T , com $T \in D$. Uma regra de classificação é definida quando identificamos um conjunto de registros, onde os valores dos atributos – $A = \{a_1, a_2, a_3, \dots a_n\}$ – levam a uma classe B , ou seja, verifica-se a ocorrência de $A \rightarrow B$ (Agrawal, Imielinski & Swami, 1993), (Hipp et al, 2002). Um modelo de aprendizagem baseado em regras é composto por expressões no seguinte formato:

SE [condição] ENTÃO [conclusão]

Por exemplo, a regra $R1$:

$R1$: SE *idade_cliente* = 25 E *profissao* = *analista_sistemas* ENTÃO *financia_veiculo_0_km* = *SIM*

As sentenças que aparecem juntamente com a indicação “SE” são conhecidas como “regras antecedentes” ou “pré-condições” e a sentença que aparece após o indicador “ENTÃO” são conhecidos como regras conseqüentes. A regra antecedente pode ser composta por um ou mais atributos de um modelo de dados usado como base de aprendizagem e respectivo valor do atributo. No exemplo citado, a condição da regra $R1$ é composta de dois atributos: a idade do cliente e a profissão; a conclusão da regra $R1$ contém a classe de predição ou conclusão da regra onde, com base nas duas condições iniciais, define-se que o cliente está apto para o financiamento de um veículo novo. A regra $R1$, pode ser ainda representada da seguinte forma:

$R1$: (*idade_cliente* = 25) $\wedge$ (*profissao* = *analista_sistemas*) $\rightarrow$ (*financia_veiculo_0_km* = *SIM*).

Se a condição composta pelas regras antecedentes são verdadeiras ou levam à conclusão, pode-se dizer que as regras antecedentes são abrangentes em relação à regra conseqüente (Tan, Steinbach & Kumar, 2005).

A qualidade de uma regra pode ser avaliada usando medidas como “cobertura” e “precisão”:

$$\text{Cobertura}(R) = |A| / |D|$$

$$\text{Precisão}(R) = |A \cap y| / |A|$$

Onde D representa um conjunto de registros ou amostra, e uma regra de classificação $R: A \rightarrow y$, a cobertura de R é definida pela fração de registros na amostra que satisfazem a sentença de pré-condição. A precisão ou fator de confiança é definida com a fração de registros onde foi satisfeita a regra R.

Um modelo de classificação baseado em regras permite a atribuição de uma classe a um registro de teste baseado na regra acionada pelo registro, ou seja, a regra que contém na sentença de pré-condição os atributos equivalentes apresentados pelo registro. Um bom modelo de classificação apresenta duas propriedades importantes:

- Regras mutuamente exclusivas: as regras definidas para um conjunto de registros são mutuamente exclusivas quando se garante que duas regras não serão acionadas pelo mesmo registro, essa propriedade deve garantir que todo registro de uma amostra seja coberto por uma única regra.
- Regras exaustivas: um conjunto de regras tem cobertura exaustiva se existir uma regra para cada combinação de atributos. Esta propriedade visa garantir que todo registro seja coberto por pelo menos uma regra dentro do conjunto de regras disponíveis.

Quando um conjunto de regras não é exaustivo, ou seja, nem todas as regras foram destacadas para classificar uma amostra, então uma regra padrão,

Rp: () $\rightarrow$ yp, deve ser incluída ao conjunto de regras. A regra padrão é atribuída quando todas as demais regras não podem ser utilizadas para classificar um registro, dessa forma, a classe padrão yp será atribuída por insuficiência de classes a serem designadas a partir das regras existentes (Han & Kamber, 2006).

Quando mais de uma regra é acionada pelo mesmo registro, há um problema de conflito na identificação de regras, que precisará ser resolvido, pois o conjunto de regras não é mutuamente exclusivo. Para a resolução desse tipo de problema, pode-se optar por duas abordagens:

- **Lista de decisão:** nesta abordagem, os elementos de um modelo baseado em regras, são organizados em ordem decrescente de acordo com sua relevância – que pode ser definida pela cobertura, precisão, ou pela ordem em que as regras são geradas. Desta forma, qualquer registro a ser classificado por regras ordenadas não será atribuído a mais de uma classe, evitando o conflito na predição de classes.

- **Regras não ordenadas:** nesta abordagem, aceita-se que um registro acione mais de uma classe, criando uma pontuação das classes que foram acionadas pelos registros. A pontuação é apurada e o registro é atribuído para a classe que recebe maior número de votos. Os modelos construídos sob essa abordagem tendem a um esforço menor na manutenção das bases de conhecimento, uma vez que não será necessário manter um procedimento para ordenação das regras. Por outro lado, a tarefa de classificação de um registro nesse tipo de abordagem exige uma capacidade de processamento computacional compatível, pois os atributos do registro de teste devem ser confrontados com todas as pré-condições das regras disponíveis.

A construção de modelos de classificação baseados em regras envolve a extração e identificação dos relacionamentos entre os atributos de um conjunto de dados e suas respectivas classes. Para tal tarefa, pode-se utilizar duas metodologias diferentes: métodos diretos, em que a extração das regras é realizada diretamente a partir da base de dados e os métodos indiretos, onde as

regras são extraídas a partir de outros modelos de classificação, tais como árvores de decisão ou redes neurais (Han & Kamber, 2006).

A apuração do número de vezes em que as combinações de atributos aparecem nas transações de dados é importante para verificação da frequência de certos valores de atributos, mas por si só não determina uma regra. As tabelas de ocorrência fornecem informações sobre as combinações que ocorrem com maior frequência nas transações de dados.

3.4.1 - Métodos diretos para extração de regras

O algoritmo de cobertura seqüencial é freqüentemente utilizado para extrações de regras diretamente da base de dados e essas são reforçadas a partir de medidas de avaliação – tais como cobertura e precisão. O critério para decidir qual regra deve ser extraída primeiro, para a composição do modelo, depende da fração de registros de treinamento que pertencem a uma classe. Basicamente, o algoritmo de cobertura sequencial começa com uma lista de decisão vazia que é preenchida à medida que uma função de extração de regras identifica a melhor regra para a classe Y e que cobre o conjunto de registros disponível na amostra de treinamento. Durante a extração das regras, todos os registros de treinamento atribuídos corretamente à classe Y são considerados classificações positivas, e os que pertencem a outras classes são contados como classificações negativas. Uma regra bem definida se destaca quando obtém quantidade considerável de classificações positivas e poucas ou nenhuma classificação negativa. Uma vez que a regra é encontrada, esta será adicionada ao final da lista de decisão, e os registros que foram cobertos pela mesma são eliminados até que todos os registros disponíveis tenham sido utilizados (Tan, Steinbach & Kumar, 2005).

Na literatura mais recente, diversos algoritmos de cobertura sequencial para extração de regras de classificação podem ser encontrados, dentre os quais pode-se citar o AQ, CN2 e o mais recente, RIPPER. Todos estes algoritmos mantêm uma estratégia em comum: as regras são identificadas uma

de cada vez e cada nova regra exclui os registros responsáveis pela sua identificação, até que o processo exclua todos os registros da amostra. A aprendizagem seqüencial de regras contrasta com a indução por árvore de decisão, pois nessa segunda abordagem, cada nó terminal ou folha representa uma regra, dessa forma, pode-se considerar que a árvore de decisão gera um conjunto de regras simultaneamente (Han & Kamber, 2006), (Michalski et al, 1986),(Cohen, 1995), (Clark & Boswell, 1991),(Clark & Niblett, 1989).

Para a aprendizagem de uma regra, considera-se a sua forma mais geral, ou seja, para uma determinada classe, considera-se o antecedente um conjunto vazio.

IF { } THEN classe = valor_classe

Numa etapa seguinte, inicia-se a avaliação dos candidatos ao termo antecedente da regra, que será composto por um conjunto de pares atributo-valor. Normalmente, os registros da amostra de treinamento podem conter vários atributos, e cada atributo pode assumir uma infinidade de possíveis valores e, nesse caso, identificar uma regra nessas condições pode se tornar computacionalmente inviável. Desta forma, uma estratégia adotada na extração de regras é que o termo consequente que melhora a qualidade da regra seja assumido, ou seja, a cada momento em que ocorre a adição de um novo atributo, deve-se fazê-lo com base em medidas de avaliação – tais como cobertura e precisão. Cada inclusão de um atributo ao termo consequente da regra deve cobrir uma quantidade maior de registros pertencentes à classe definida.

Para demonstrar o processo de extração de regras, toma-se como exemplo o processo de extração de regras para classificar clientes com base nas aquisições de financiamento para compra de veículos. Alguns atributos tais como faixa etária do cliente, sexo, estado civil, veículo adquirido, valor e idade do veículo a ser adquirido fazem parte da base estudada. A estratégia é identificar regras para a concessão de crédito para financiamento de veículo ou negociação de dívida.

Na primeira iteração, a regra não possui o termo antecedente definido para a classe crédito_concedido = 'SIM'. Com base na precisão da classe, o primeiro atributo destacado na base de teste refere-se à faixa etária do cliente:

IF *faixa_etaria_50_57* = "sim" ENTAO *credito_concedido* = "SIM".

Em seguida, a regra pode ser expandida com a inclusão de um novo atributo referente ao modelo de veículo financiado:

IF *faixa_etaria_50_57* = sim" AND *modelo_veiculo* = "ASTRA" ENTAO *credito_concedido* = "SIM".

3.5 - Inferência Probabilística para Tratamento de Incertezas em Modelos de Classificação

A inferência probabilística é uma das ferramentas mais utilizadas para minimizar a incerteza presente nas bases de conhecimento. Nos projetos de sistemas de apoio à decisão, um fator importante que deve ser considerado é o porte da base de conhecimento. Isto porque a inferência probabilística, utilizando como base o Teorema de Bayes, pode apresentar problemas de performance, considerando a grande matriz de probabilidades gerada pela quantidade de evidências presentes nos cenários. Por outro lado, deve-se considerar que um sistema de apoio à decisão, implementado para realizar classificações ou predições, terá melhor desempenho se possuir um registro que mapeie a maior quantidade possível de experiências – as lições aprendidas. Com isso, as questões de “banco de dados vazio” ou uma condição oposta podem afetar a precisão de suas saídas, levando possivelmente a uma queda de performance. Em algumas situações onde a modelagem da base de conhecimento expõe uma grande quantidade de atributos, pode-se optar por descartar os atributos menos relevantes, avaliando-se que a ausência destes ainda mantém o modelo dentro de parâmetros confiáveis.

Como forma de contornar os problemas citados acima, os primeiros sistemas de apoio à decisão, desenvolvidos nos anos 60, tratavam a incerteza

de forma restritiva (Luger, 2004). Por exemplo, sistemas de apoio de diagnóstico de enfermidades, desenvolvidos usando essa estratégia, assumiam um conjunto de possíveis doenças a serem diagnosticadas e que estas eram mutuamente excludentes, sendo que a evidência fosse condicionalmente independente dada qualquer hipótese. Assim, somente uma doença poderia ser assumida para ser diagnosticada para cada paciente. O resultado é que os sistemas implementados, usando essa abordagem, inferiam sobre um conjunto com pequeno número de hipóteses e evidências limitadas. Por essa razão, o interesse, naquela época, de se utilizar a inferência probabilística diminuiu por se imaginar que esta seria inadequada para expressar a estrutura do conhecimento humano. Adicionalmente, acreditava-se que se fosse aplicada a domínios maiores, as simplificações adotadas produziriam resultados incorretos e as conclusões a que esses sistemas chegassem não seriam confiáveis para os especialistas do domínio.

Na década de 80, novamente observou-se uma motivação para a utilização dessa abordagem. Porém, desta vez com a consideração do relacionamento causal e a independência condicional entre variáveis do domínio. Neste caso, seria necessário representar as probabilidades condicionais somente entre variáveis diretamente dependentes, e com isso, viabilizaria a implementação de aplicações utilizando a inferência probabilística. Esta retomada se deveu às pesquisas direcionadas em modelos baseados em representações gráficas denominadas *redes probabilísticas bayesianas*, e através das quais permitem representar e manipular a incerteza com base em princípios matemáticos fundamentados e modelar o conhecimento do especialista do domínio de forma intuitiva (Langley, 1992), .

A utilização da representação gráfica das redes probabilísticas permite explicitar as relações de dependências, tornando-se uma ferramenta poderosa na aquisição de conhecimento e no processo de validação do modelo gerado. Desta forma, existindo uma representação do domínio sob a forma de rede causal, se houver uma evidência, é possível determinar quais as variáveis que devem ter a razão de verossimilhança atualizada (Jensen, 1990).

3.6 - Classificadores Bayesianos – *Naïve Bayes*

O classificador *Naïve Bayes* é considerado um dos classificadores Bayesianos mais simples, sendo naturalmente incremental devido à sua simplicidade (Yang, 2003),(Alcobé, 2002),(Keogh & Pazzani, 1999). Dessa forma, ele pode ser aplicado a uma gama diversificada de problemas com resultados satisfatórios se comparados a esquemas de aprendizagem mais sofisticados. (Duda,Hart & Stork, 2001),(Langrey, Iba, Thompson, 1992), (Ramoni & Sebastini, 2001), (Domingos & Pazzani, 1997). O classificador *Naïve Bayes* considera que o efeito do valor de um atributo em uma determinada classe é independente dos valores dos demais atributos e dessa forma, o classificador *Naïve Bayes* é conhecido por assumir a independência condicional (Lewis, 1998).

3.6.1 - Classificação de eventos utilizando o Teorema de Bayes

Considerando um conjunto de classes C , ω_1 , ω_2 , ... ω_C , com probabilidades conhecidas *a priori* para cada classe de $p(\omega_1)$, $p(\omega_2)$,... $p(\omega_C)$. Supondo que se observa um vetor de medida x , e se deseja atribuir esse vetor x a uma das classes C . A regra de decisão baseada em probabilidades é a atribuição de x a uma classe ω_j , se a probabilidade da classe ω_j dada a observação de x , $p(\omega_j|x)$, é maior que as probabilidades de todas as outras classes ω_1 , ω_2 , ... ω_C , dado x . Ou seja, atribui-se x à classe ω_j se:

$$p(\omega_j|x) > p(\omega_k|x) \quad k = 1,...,C; k \neq j \quad (3.1)$$

A probabilidade a posteriori $p(\omega_j|x)$ pode ser expressa em termos da probabilidade a priori e da função de densidade condicional da classe $p(x|\omega_j)$ usando o teorema de Bayes, como:

$$p(\omega_i|x) = \frac{p(x|\omega_i) p(\omega_i)}{p(x)}$$

A regra de decisão descrita em (3.1), para a atribuição do vetor x à classe ω_j pode ser escrita como:

$$p(x|\omega_j) p(\omega_j) > p(x|\omega_k) p(\omega_k) \quad k = 1, \dots, C; \quad k \neq j \quad (3.2)$$

Para duas classes, a regra de decisão (3.2) pode ser escrita como:

$$l_r(x) = \frac{p(x|\omega_j)}{p(x|\omega_i)} > \frac{p(\omega_j)}{p(\omega_i)}$$

implica que x pertence à classe ω_j onde $l_r(x)$ é a função de verossimilhança.

3.7 - Pós- Processamento do Processo KDD

A etapa de pós-processamento dos modelos KDD abordados neste trabalho, definem dois objetivos distintos após a execução das atividades de mineração de dados:

- 1) A avaliação de desempenho do modelo de aprendizagem;
- 2) A disponibilização da base de conhecimento.

A avaliação de desempenho é uma etapa praticamente implícita na fase de mineração de dados, pois é essencial que um modelo gerado por técnicas de mineração de dados seja de alguma forma confrontado por procedimentos sistemáticos para verificar seu grau de precisão a partir de uma base de teste. Algumas ferramentas que implementam algoritmos de

classificação, por exemplo Weka, permitem a escolha do procedimento de teste pelo usuário antes mesmo de executar o procedimento de geração do modelo de aprendizagem (Han & Kamber, 2006).

A disponibilização da base de conhecimento é uma tarefa de homologação e preparação da estrutura física da base de conhecimento para que possa ser utilizada a partir de ferramentas gráficas ou então, alimentando-se uma base de dados, com tabelas que representam os indicadores do modelo de aprendizagem, tais como os valores das probabilidades condicionais de um atributo para as classes estudadas.

A avaliação do desempenho do modelo de classificação pode ser feita pela contagem e tabulação dos registros classificados corretamente e incorretamente, essa tabela é conhecida como *matriz de confusão* (Tan, Steinbach & Kumar, 2005). A tabela 3.1 mostra as predições feitas corretamente para as classe que contém duas classes e as quantidades de predições para as mesmas classes atribuídas incorretamente.

Tabela 3.1: Avaliação do modelo de classificação.

Avaliação de Performance Do Modelo		Atribuição por classificação	
		Classe_1	Classe_2
Classe Real	Classe_1	f_{11}	f_{21}
	Classe_2	f_{12}	f_{22}

Na tabela 3.1, as quantidades expressas nos campos f11 e f22 foram classificadas corretamente, enquanto que o valor expresso por f12 representa a quantidade de registros atribuídos à classe 1 quando na realidade deveriam pertencer à classe 2 e vice-versa para o valor expresso por f21. Para se ter uma métrica de performance do modelo de classificação, utiliza-se a seguinte fórmula:

$$Precisao = \frac{Qtd_Classificações_Corretas}{Qtd_Total_Classificações} = \frac{f11 + f22}{f11 + f12 + f21 + f22}$$

De forma equivalente, pode-se calcular a *taxa de erro*, considerando, no numerador, a quantidade de classificações realizadas incorretamente.

$$Erro = \frac{f12 + f21}{f11 + f12 + f21 + f22}$$

Esse procedimento permite a apuração dos resultados de um modelo de classificação perante uma base de teste contendo as classes reais. O modelo de classificação deverá atribuir um valor a cada registro da base de teste e ao final do processo de classificação, as classes atribuídas com o uso do modelo serão comparadas às classes reais e os resultados serão computados na matriz de apuração do modelo.

4. Aplicações do Processo KDD

Este capítulo apresenta os resultados obtidos na resolução dos problemas encontrados em empresas que necessitam de recursos de suporte à decisão para ações de vendas que promovam a retenção, aquisição e vendas concorrentes de produtos às suas carteiras de clientes. O maior problema identificado nessas empresas é a composição de grupos-alvo de ações de marketing e divulgação de produtos, contextualizando a implementação de processos de extração de conhecimento em banco de dados. A motivação da proposta resultante da pesquisa neste trabalho está orientada à possibilidade da obtenção de conhecimento a partir de abordagens que otimizam a exploração dos dados nas empresas para incentivar o investimento em tecnologias, dotando-as com recursos que permitam controlar todo o fluxo da informação. Como resultado imediato desta pesquisa, espera-se uma melhoria significativa dos processos internos, reduzindo custos e tornando a cadeia produtiva ajustável e flexível às novas necessidades de seus produtos e clientes.

4.1 - Estudo de Retenção de Clientes

O conceito de retenção de clientes na área de marketing visa estabelecer um conjunto de metas para garantir a sobrevivência de um grupo de clientes ao longo de um período (Berry & Linoff, 2004). O primeiro elemento a ser considerado no estudo de retenção de clientes é o período de tempo em que se deseja manter um grupo de clientes como consumidores de um produto ou serviço e, nesse item, destaca-se o primeiro problema que o estudo de retenção deve solucionar. Em alguns ramos de atuação, os pontos que podem ser considerados como “início” e “fim” de um relacionamento consumidor-produto precisam ser estabelecidos. Inicialmente, pode-se supor que o conceito início-fim do relacionamento entre um cliente e uma empresa é algo trivial. Analisando os produtos ou serviços que mantêm datas limites em seus contratos, dentro dos quais, empresa e clientes possuem responsabilidades e direitos, pode-se considerar essas datas como pontos de início e fim. Em outros ramos de atuação, as datas limites de validade de contratos não são estabelecidas ou não se aplicam ao negócio e, portanto identificar esses pontos de início-fim não é algo trivial. Por exemplo, pode-se identificar esse cenário em supermercados e lojas virtuais, onde a entrega do produto da última compra efetuada pelo cliente pode ou não ser observada como o final do relacionamento entre o consumidor e a empresa. Cada estabelecimento pode definí-la de forma diferenciada, por exemplo, considerando o final da relação se o cliente não fizer um novo pedido dentro de um determinado período.

A retenção de clientes pode ser acompanhada a partir da representação gráfica da proporção de clientes retidos em um determinado período de tempo, mais conhecido como curva de retenção. A curva de retenção é um tipo de histograma acumulativo, uma vez que um cliente retido no terceiro ano após o início da apuração também esteve presente nas apurações dos dois anos anteriores. Para se ter uma visualização do desempenho entre carteiras de clientes, a representação gráfica pode expor o comportamento de vários grupos de clientes. A figura 4.1 exhibe a curva de retenção de três grupos de clientes,

apurando a sua permanência a cada 12 meses. Inicialmente, adotou-se o valor 1 no eixo dos valores por representar a carteira com todos os clientes. Ao longo de 36 meses, em todos os grupos analisados, a evasão maior nos três grupos ocorre nos primeiros 12 meses e consideradas as perdas ao final de 24 meses, cada grupo possui menos de 50% dos clientes. Os registros analisados são de clientes que possuem contrato de seguro anual, sendo que em cada renovação, e desde que não ocorra sinistro durante a vigência do contrato anterior, será aplicado um desconto ao valor do prêmio pela fidelidade do cliente.

Figura 4.1: Curva de retenção de clientes.

4.1.1 - Metodologia aplicada ao Pré-Processamento de dados

O estudo de retenção de clientes foi organizado nas três etapas do processo KDD proposto neste trabalho, organizando as atividades do projeto em três etapas: pré-processamento de dados, mineração de dados e pós-processamento. Na etapa de pré-processamento, as atividades foram executadas conforme a sequência detalhada na tabela 4.1, tendo como produto final as bases de treinamento e teste dos modelos de aprendizagem.

Tabela 4.1: Atividades da etapa de pré-processamento do projeto de retenção de clientes.

Sequencia	Atividade	Descrição
1	Objetivo do projeto	Objetivo do projeto KDD, delimitar o problema
2	Análise do processo	Estudar o processo organizacional - origem dos dados
3	Análise do modelo de dados	Estudar o modelo de dados mantido pelos sistemas transacionais ou Data Warehouse
4	Avaliação atributos indisponíveis	Verificar as informações desejáveis na base de dados
5	Avaliação de atributos gerados por transformação	Identificar campos para transformação e padronização
6	Avaliação de dados <i>missing</i>	Identificar volume de dados missing
7	Extração de dados	Implementar rotina para extração ou convergência de bases
8	Limpeza	Executar processo de limpeza de dados inválidos

A retenção de clientes foi abordada através da criação de um modelo de suporte à decisão para uma análise preditiva de registros de clientes como forma de identificar clientes propensos a não renovarem os seus contratos. Desta forma, iniciou-se um estudo de exploração de dados, visando verificar a influência das alterações contratuais nos contratos vigentes como elemento que leva ao abandono no período seguinte. Face ao problema apresentado, identificou-se que a análise de dados tem como objetivo classificar os clientes quanto à expectativa de sua permanência e, dessa forma, o estudo teve como objetivo gerar modelos descritivos do comportamento de permanência ou abandono.

A análise de dados da fase de pré-processamento envolveu a integração de dados, combinando dados de quatro fontes distintas, entre bancos de dados de sistemas transacionais e data warehouse. Durante a fase de integração de dados, a atividade de avaliação dos modelos de dados tornou-se implícita, uma vez que alguns atributos que compartilhavam do mesmo conceito foram criados nos modelos de dados com nomes e tipos de dados diferentes.

Desta forma, é imprescindível o envolvimento de um analista de sistemas que tenha grande familiaridade com os conceitos de negócio e a forma como tais conceitos foram interpretados para a criação de um modelo de dados relacional, pois atributos que conceitualmente representam o número de identificação do cliente, ao serem criados em tabelas de bancos de dados diferentes, assumiram outros nomes físicos que dificultaram sua identificação no modelo lógico, tais como “*clicod*” ou “*custid*”. Este problema e sua resolução, através da criação de um nome de atributo padrão, também é conhecido como *problema de identificação de entidades* (Han & Kamber, 2006). A solução para este problema envolve a identificação de todos os conceitos lógicos, atributos e tipos de dados e sua aplicação ao modelo físico que, neste projeto implicou na implementação de um programa específico para a transformação de dados. A tabela 4.2 apresenta alguns atributos lógicos e a padronização aplicada aos valores aceitos para cada campo.

Tabela 4.2: Mapeamento dos atributos lógicos e campos físicos do modelo de dados.

Conceito - descrição	Nome Lógico	Nome Físico
Código do cliente – código numérico único que identifica um cliente	Cliente_ID	Clicod
Tipo de pessoa – sigla que identifica se o cliente é pessoa física ou jurídica	Cliente_Tipo_Pes	Tippes
Sexo do cliente – código identificador do sexo do cliente	Cliente_Sexo_Sgl	Sglsex
Estado civil – código identificador do estado civil do cliente	Cliente_Estado_Civil	Codestcvl
Situação do cliente – código identificador da situação do relacionamento entre o cliente e a empresa	Cliente_Situação	Sttcliente
Data de entrada do cliente – data de início do primeiro contrato entre cliente e empresa	Data_Inicio_Relacionamento	Datincrel

Outro aspecto também avaliado, com respeito às transformações de dados para padronização conceitual, refere-se aos valores assumidos pelos atributos, sendo realizados em dois níveis: mapeamento 1 para 1 e redução.

No mapeamento 1 para 1 mantém-se um código único por valor, uma vez que as bases de dados podem utilizar diferentes grupos de valores para representar os seus atributos. No atributo “estado civil”, por exemplo, o valor “solteiro” pode ser representado com o carácter “S” em um dos repositórios de dados e em outra fonte, encontra-se o mesmo conceito aceitando o número “1” para indicar o “estado civil = solteiro”. Nesse mapeamento, todos os valores possuem um valor padrão equivalente que o representa.

Na criação dos códigos ou possíveis valores de um atributo categórico, considera-se o valor que mantém unicidade conceitual e, dessa forma, não possui outro equivalente que possa ser utilizado para representar a informação desejada. Porém, nem sempre o mapeamento conceitual 1 para 1 é necessário. Alguns tipos de problema determinam que os valores podem ser novamente agrupados, formando uma nova lista de valores possíveis a um atributo. Esse tratamento de dados pode ser considerado como um tipo de redução. No problema abordado neste trabalho, o comportamento dos grupos analisados, considerando-se o estado civil, não necessita que os agrupamentos sejam realizados em torno de cada um dos possíveis valores do atributo: o que se pretende analisar é o comportamento do grupo baseado na condição de cliente possuir uma relação civil estável. Sendo assim, de acordo com o problema, solteiros e divorciados são parte de um mesmo grupo por não possuírem relação civil estável. A redução conceitual aplicada aos valores dos atributos parte de um entendimento claro do problema e os novos agrupamentos de valores não devem ser refletidos nas bases de origem, pois, neste caso, identificou-se uma redundância conceitual a partir do contexto do problema, mas em linhas gerais, “solteiros” e “divorciados” são condições diferentes de estado civil.

A fase de pré-processamento foi concluída com a extração das bases de treinamento e teste. Esta etapa envolveu a construção de um programa para geração de arquivo com dados unificados, considerando alguns atributos já conhecidos em função dos modelos de retenção gerados neste trabalho. Basicamente, o programa de unificação de bases de dados apresenta uma estrutura básica contemplando as seguintes elementos: (1) unidade de extração de dados, (2) unidade de transformação e eventualmente (3) unidade de agregação, conforme figura 4.3.

Figura 4.2: Fluxo de extração de dados de um programa ETL para projeto KDD.

Neste trabalho, fica evidenciado que a documentação do fluxo de extração de dados deve ser mantida como forma de preservar e controlar as versões dos dados da extração e, ainda, diminuir o tempo gasto na incorporação de novos atributos. A figura 4.2 apresenta um fluxo básico, com o mínimo de elementos ou módulos, que torna-se complexo na medida que novos acessos a outras fontes de dados são incluídas, novas unidades de tratamentos de valores, alterações nas rotinas limpeza e resolução das multiplicidades de registros, conforme a granularidade dos objetos de dados extraídos. Abaixo, o objetivo de cada elemento do fluxo de extração é descrito brevemente:

(1) Unidade de extração de dados: tem como objetivo encapsular as configurações de conexão aos bancos de dados ou outros repositórios de dados, tabelas que serão acessadas, campos e alguns critérios de inclusão de registros – também conhecidos como “*where conditions*”.

(2) Unidade de transformação: atua na transição dos dados, aplicando regras de limpeza, correção de divergências, filtros de exclusão, dentre outros.

(3) Unidade de agregação: tem como objetivo solucionar os problemas de multiplicidade de registros em função de extrações de dados com diferentes níveis de granularidade.

A etapa de pré-processamento foi encerrada com a extração dos arquivos de treinamento e teste em formato de texto. Optou-se pela geração de arquivos para que as amostras das bases de treinamento e teste fossem preservadas de ações de manutenção das bases de dados. Do ponto de vista de acesso às bases de dados, a criação de um ambiente isolado para os projetos de mineração de dados preserva o sistema dos acessos concorrentes.

4.1.2 - Modelo de retenção com subclasse fixa

A etapa de mineração de dados iniciou-se com a preparação do ambiente de análise de dados e com base no problema apresentado, realizou-se a escolha da técnica de mineração de dados. O ambiente de mineração de dados foi isolado do ambiente de *data warehouse* para preservar o sistema de múltiplos acessos e controle do histórico de dados.

O primeiro modelo gerado foi elaborado a partir de 144 registros de clientes que alteraram seus contratos. Os atributos disponíveis na base analisada foram distribuídos conforme as tabelas 4.3 a 4.6, os quais contemplam as probabilidades de cada atributo sobre a classe “Retenção”. A tabela 4.7 contém uma distribuição geral quanto à classe de retenção, considerando a amostra de treinamento do Modelo 1. Durante a execução da atividade de extração da base de dados, foi identificada uma grande quantidade de registros incompletos, os *missing values*, o que resultou na exclusão de 10% dos registros do modelo 2

Tabela 4.3: Posições das retenções em relação a alteração do produto.

ALTERACAO PRODUTO	MARCA	RETENCAO	
		SIM	NÃO
SIM		0,687	0,656
NÃO		0,313	0,344
TOTAL		1	1

Tabela 4.4: Posições das retenções em relação valor contratado.

AUMENTOU VALOR	RETENCAO	
	SIM	NÃO
SIM	0,857	0,875
NÃO	0,143	0,125
TOTAL	1	1

Tabela 4.5: Posições das retenções em relação ao mês em que ocorreu a alteração.

MÊS DE ALTERACAO	RETENCAO	
	SIM	NÃO
1 ao 4	0,348	0,219
5 ao 8	0,375	0,406
9 ao 12	0,277	0,375
TOTAL	1	1

Tabela 4.6: Posições das retenções em relação ao ano em que iniciou a relação cliente-empresa.

ANO REFERÊNCIA	RETENCAO	
	SIM	NÃO
X < 2000	0,054	0,031
2001 – 2004	0,134	0,125
2005 – 2008	0,804	0,844
TOTAL	1	1

Tabela 4.7: Probabilidade de retenção da amostra de treinamento da primeira iteração.

RETENCAO	
SIM	NÃO
0,778	0,222

Para a realização das classificações, criou-se um ambiente para o processamento de lotes de registros na ferramenta "Access". A tabela central – com os registros de teste foi relacionada às tabelas do modelo, conforme a figura 4.3.

Figura 4.3: Entidades do modelo de aprendizagem associada à base de teste.

Para as classificações de uma base de teste, realizou-se consultas de acordo com a quantidade de classes. No estudo realizado, a classe de retenção poderia assumir dois valores, portanto, o arquivo de teste foi usado para a realização de dois acessos, cada qual retornando as respectivas probabilidades em relação à classe fixada.

Para cada $C\{c_1, \dots, c_n\}$ FAÇA

Entrada: $X = \{x_1 = a_1, x_2 = a_2, x_3 = a_3, x_4 = a_4, \dots, x_n = a_n, C = c\}$

Para cada $[x = a, C = c]$ com $x = 1, \dots, n$ FAÇA

SELECIONA prb_x_valor DA TABELA $Probabilidade_x$

COM $x = a$

E classe = c

Saida: $P = \{p(x_1 = a_1 | c), p(x_2 = a_2 | c), \dots, p(x_n = a_m | c)\}$

FIM

Os dados da base de teste, associados às probabilidades, foram alocados em planilha, onde foi criada uma rotina com base na técnica Naive Bayes, conforme figura 4.4.

Atributos Classe							Probabilidades Condicionais e Prob. Da Classe					
r1	x1	x2	x3		xn	Y	$p(x_1 y=1)$	$p(x_2 y=1)$	$p(x_3 y=1)$	...	$p(x_n y=1)$	$p(y=1)$ $P(y=1 X)$
							$p(x_1 y=0)$	$p(x_2 y=0)$	$p(x_3 y=0)$	...	$p(x_n y=0)$	$p(y=0)$ $P(y=0 X)$

Figura 4.4: Estrutura da planilha de classificação de registros.

Para a classificação dos registros:

SE $P(y = 1 | X) > P(y = 0 | X)$ ENTÃO

$Y_p = 1$

Caso contrário

$Y_p = 0$

FIM_SE

Cada valor de classe identificada na rotina de classificação foi alocada em coluna, para apuração dos resultados. Dessa forma, se a classe do registro Y for igual a Y_p , a classificação foi realizada corretamente e lançado o valor “1” ao indicador de desempenho do classificador nas classificações corretas e “0” nas classificações incorretas, conforme figura 4.5.

Estrutura de Apuração		
Y	Yp	Indicador
1	1	1
1	1	1
1	0	0
0	1	0
0	0	1
1	1	1

Classe do registro de teste →
 ↑
Valor definido pelo classificador

Figura 4.5: Estrutura de apuração de resultados das classificações.

Na primeira iteração de teste do Modelo 1, foi realizada a classificação de uma amostra contendo 100 registros, obtendo-se a configuração de resultados conforme demonstrado na tabela 4.8. A primeira interação mostra um resultado muito comum quando o método de seleção de dados não prevê a criação de amostras balanceadas, com classes positivas e negativas. Este procedimento foi observado na empresa avaliada, onde as análises de dados ações de vendas eram executadas diretamente de uma amostra sem considerar o problema das *amostras localizadas* (Han & Kamber, 2006).

Tabela 4.8: Distribuição de resultados na primeira amostra de teste para o modelo 1.

Classe de Teste	Classe de Predição	
	RET = SIM	RET = NÃO
RET = SIM	59	0
RET = NÃO	41	0

Contabilizando os resultados de precisão (P) e taxa de erro (Er), conforme as equações abaixo:

$$P = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

e

$$Er = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

o modelo 1 possui uma precisão $P = 0,59$ e $Er = 0,41$. Para um modelo inicial, a primeira etapa de teste demonstrou que embora o modelo tenha uma precisão considerável para uma amostra pequena e que não tenha gerado falsos negativos, o modelo não conseguiu identificar os casos de abandono, quando a classe de teste "RET = Não", ou seja, classificou 41 registros de abandono como clientes que não apresentariam risco de perda. Desta forma, como o objetivo era prever o risco de abandono por parte do cliente, o modelo falharia no seu objetivo essencial.

No contexto do volume de dados para o qual o modelo 1 será aplicado, verificou-se que o mesmo passaria a apresentar problemas nas classificações futuras, uma vez que uma base de testes contendo 144 registros poderia exibir uma tendência que não seria necessariamente a tendência de uma base maior. Elaborar um modelo com uma amostra pequena pode ser útil quando as classes estão representadas proporcionalmente em relação a uma amostra maior – por exemplo, as apurações contratuais realizadas mensalmente, semestralmente, etc. Por outro lado, isso não é uma garantia de que os valores dos atributos também estarão representados na mesma proporção em relação a uma base com maior quantidade de elementos. É possível visualizar padrões em uma amostra pequena se o espaço de valores que podem ser assumidos pelos atributos é mínimo, por exemplo, algumas transformações de dados aplicadas na fase de pré-processamento relacionam os valores dos campos de uma tabela a valores binários.

A utilização das amostras parciais muitas vezes é premissa de um problema em que a análise se restringe a grupos específicos de dados a partir dos quais serão gerados vários modelos. Por exemplo, ao se criar um modelo descritivo com o comportamento do consumo de energia, provavelmente haverá comportamentos diferentes de acordo com os dias da semana e as regiões onde se concentra o consumo. Um modelo que expressa o comportamento de consumo aos sábados pela manhã pode não ser o melhor modelo para descrever o consumo de energia nos demais dias da semana, mesmo fixando uma faixa de horário equivalente. Dessa forma, é possível a geração de submodelos descritivos derivados de um modelo de aprendizagem inicial para exibir cenários mais específicos, por exemplo, os modelos descritivos do fluxo de veículos podem ser inicialmente gerados para cobrir os dias da semana para planejamento de rotas alternativas para melhoria da fluidez e posteriormente, conforme o desempenho do modelo geral seja necessário criar um modelo mais específico abrangendo períodos ou faixas de horário.

O segundo modelo de aprendizagem, gerado com outra base de treinamento e desta vez contendo 56.000 registros, apresentou a configuração de probabilidades de retenção conforme a tabela 4.9. Os atributos analisados foram os mesmos do modelo 1. Comparado ao modelo 1, o índice de retenção e abandono do modelo 2, 0,7309 e 0,269 respectivamente, ajustou os valores aplicados no cálculo das probabilidades de classificações positivas e negativas cada registro, contribuindo para a melhoria do classificador em relação às classes negativas.

Tabela 4.9: Distribuição de probabilidades de retenção do modelo 2.

ATRIBUTO – VALOR	RETENÇÃO	
	P_SIM	P_NÃO
AUMENTOU VALOR = SIM	0,919	0,844
AUMENTOU VALOR = NÃO	0,081	0,156
ALTEROU PRODUTO = SIM	0,609	0,630
ALTEROU PRODUTO= NÃO	0,391	0,370
MÊS CONTRATO =0	0,019	0,014
MÊS CONTRATO = 1	0,073	0,054
MÊS CONTRATO = 2	0,082	0,061
MÊS CONTRATO = 3	0,089	0,074
MÊS CONTRATO = 4	0,088	0,084
MÊS CONTRATO = 5	0,091	0,088
MÊS CONTRATO = 6	0,095	0,092
MÊS CONTRATO = 7	0,095	0,105
MÊS CONTRATO = 8	0,096	0,109
MÊS CONTRATO = 9	0,094	0,109
MÊS CONTRATO = 10	0,091	0,107
MÊS CONTRATO = 11	0,075	0,090
MÊS CONTRATO = 12	0,013	0,013
ANO INICIO CONTRATO > 2000	0,064	0,087
ANO INICIO CONTRATO: 2001 - 2004	0,176	0,218
ANO INICIO CONTRATO : 2005 - 2008	0,76	0,695

A matriz de contagem das classificações gerada do modelo 2, após a classificação de uma massa de teste composta por 100 registros, é descrita na tabela 4.10.

Tabela 4.10: Resultados da massa de teste 1 com o modelo de aprendizagem 2.

Classe de Teste	Classe de Predição	
	RET = SIM	RET = NÃO
RET = SIM	30	25
RET = NÃO	33	12

Conforme os valores apresentados na tabela 4.10, se obtém os seguintes resultados para o modelo de classificação:

- 42 classificações corretas para a classe *Renovação* com os valores "SIM" e "NÃO".
- 58 classificações incorretas, sendo 33 falsos positivos e 25 falsos negativos.

Desta forma, o modelo de classificação obteve a seguinte pontuação para precisão (P) e taxa de erro (Er):

$$P = 0,42 \text{ e } Er = 0,58$$

Uma segunda massa de testes foi selecionada para a validação do modelo, desta vez utilizando 150 registros.

Tabela 4.11: Resultados da massa de teste 2 com o modelo de aprendizagem 2.

Classe de Teste	Classe de Predição	
	RET = SIM	RET = NÃO
RET = SIM	111	1
RET = NÃO	36	2

De acordo com os resultados obtidos na segunda massa de testes, os valores de precisão e taxa de erro foram os seguintes:

$$P = 0,75 \text{ e } Er = 0,25$$

Após a terceira iteração de validação do modelo, o resultado das classificações apontou que o modelo de classificação não detectava as classes negativas, como demonstrado nas tabelas 4.12 e 4.13:

Tabela 4.12: Resultados da massa de teste 3 com o modelo de aprendizagem 2.

Classe de Teste	Classe de Predição	
	RET = SIM	RET = NÃO
RET = SIM	370	0
RET = NÃO	230	0

Tabela 4.13: Resultados da massa de teste 4 com o modelo de aprendizagem 2.

Classe de Teste	Classe de Predição	
	RET = SIM	RET= NÃO
RET = SIM	1421	0
RET = NÃO	497	0

A utilização dos dois indicadores de desempenho, precisão e taxa de erro são úteis na avaliação geral do modelo 2 no contexto das classificações negativas e positivas. Uma análise dos dois resultados acima poderia levar a se considerar satisfatório o desempenho do modelo, considerando-se isoladamente a precisão e erro da classe positiva de retenção, já que a precisão da amostra 4, em relação à amostra 3, foi superior (0,74 contra 0,61) e taxas de erro inferior na base de teste 4. Por outro lado, o modelo 2 não conseguiu identificar as classes negativas, mesmo em uma amostra considerável, contendo mais de 1900 registros.

Após a avaliação com uma massa de teste contendo 10000 registros, obteve-se os seguintes resultados do modelo de classificação, com precisão $P = 0,6556$ e taxa de erro $Er = 0,3444$, conforme tabela 4.14.

Tabela 4.14: Resultados da massa de teste 5 com o modelo de aprendizagem 2.

Classe de Teste	Classe de Predição	
	RET = SIM	RET = NÃO
RET = SIM	6072	679
RET = NÃO	2765	484

Os resultados apontados com as amostras de teste 3 e 4 do modelo 2, indicam que essas amostras contém registros localizados, o que de fato foi verificado. As amostras selecionadas, mesmo com uma seleção aleatória, acabaram sendo formadas por registros com distribuição desproporcional para determinados valores de atributos da classe negativa de retenção. Este ponto foi totalmente verificado, uma vez que o próprio modelo poderia ser descartado após as etapas de teste com as amostras 3 e 4, e mesmo com amostras

maiores, como a amostra 4, o modelo não identificou classes negativas. Após a última iteração de teste com o modelo 2, verificou-se a presença das classificações negativas, mas com um fator de erro alto para essa classe, pois o modelo errou nas classificações de 2765 registro negativos, classificados como positivos.

Devido à indisponibilidade de dados ou ao desconhecimento dos atributos que compõem uma base de dados, percebe-se que muitos modelos são criados com uma quantidade menor de registros, e uma vez avaliado um desempenho considerado satisfatório, ocorrem novas iterações na geração de um modelo envolvendo uma quantidade maior de registros. Este tipo de procedimento deve ser evitado, pois um modelo que será gerado para a classificação de uma grande quantidade de registros irá provavelmente conter registros numa configuração não verificada em uma amostra menor. Uma amostra pequena pode não conter exemplos que cobrem todos os possíveis valores de um atributo ou ainda, a amostra extraída contempla os registros em que os valores dos atributos não apresentam necessariamente a mesma configuração de distribuição da amostra total ou população.

A inclusão de um atributo ao modelo pode ser um recurso para a melhoria do desempenho do modelo, mas essa evolução deve ser acompanhada por testes de apuração de classificações. Com base no modelo 2, um novo atributo, a idade do cliente, foi adicionado ao modelo de aprendizagem e avaliada uma nova massa com 10000 registros, conforme destacado na tabela 4.15. Comparado aos resultados do teste com a amostra da tabela 4.14, os resultados obtidos mostram uma melhora sensível no desempenho do modelo, pois a precisão nas classificações passou a ser $P=0,69$ e erro $Er=0,3$.

Tabela 4.15: Resultados da massa de teste 6 com o modelo de aprendizagem 2.

Classe de Teste	Classe de Predição	
	RET = SIM	RET = NÃO
RET = SIM	6558	309
RET = NÃO	2773	360

O modelo gerado nesta seção manteve uma subclasse fixa, uma vez que o modelo gerado descreve as retenções com base nas alterações contratuais. Mantendo o atributo fixo, extraiu-se da amostra original apenas os clientes que efetuaram alterações, e verificou-se o comportamento de permanência desse grupo específico de clientes. A motivação para este tipo de investigação normalmente parte do cliente do projeto, como forma de verificar as implicações de eventos ocorridos durante um contrato vigente. Uma das dificuldades geradas pelas subclasses pré-definidas é a necessidade de se obter novos atributos muitas vezes não presentes na base de dados original. Neste caso específico, necessita-se de atributos que caracterizam a relação cliente-fornecedor antes e depois da alteração contratual.

4.1.3 - Modelo de classificação descritivo para estudo de retenção

O modelo de classificação gerado para descrever uma amostra de dados em torno das retenções utilizou o fluxo de extração de dados descrito na figura 4.2. A base de dados utilizada para a extração contava com aproximadamente 280.000 registros. A aplicação do processo de extração resultou em uma base de treinamento com 65.533 registros. Durante a etapa de transformação de dados, foram aplicadas algumas regras com a finalidade de excluir dados divergentes e indisponíveis, conforme demonstrado na tabela 4.15. Além disso, aplicou-se regras de inclusão de registros, como forma de evitar a dispersão dos dados em torno de valores com pouca representação. Este critério afetou o atributo “Código_do_Produto” (COD_Produto), onde foram selecionados os produtos 1, 2, 3 e 4. Para cada registro de contrato, o programa de extração de registros identificou a classe de Retenção, atribuindo o valor “1” (“SIM”) na permanência do cliente em contrato imediatamente posterior ou “0” (“NÃO”) para os clientes que não permaneceram com a empresa após a finalização dos seus contratos.

Tabela 4.15: Regras para limpeza de dados aplicados no fluxo de extração de dados.

Regra	Campo	Parcial	Afetados	Removidos	Ação
Cliente sem idade informada	IDD_Client	69292	734	734	Exclusão
Contrato sem ano de inclusão	ANO_INC_CONT	68558	23	23	Exclusão
Idade do cliente superior a 100 anos	IDD_Client	68535	6	2	Revisão e Correção
Cliente com produto não comercializado ou especial	COD_PRODUTO	68533	3000	3000	Exclusão
Total		65533	3763	3759	

Com a geração da base de treinamento, iniciou-se o processo de cálculo das probabilidades condicionais para cada par atributo-valor em relação à classe de retenção, $P(x | y)$, sendo x a sentença composta pelo atributo e respectivo valor e y a classe associada ao registro. Neste caso, a retenção assume o valor “SIM” ou “NÃO”. O resultado dos cálculos foram incluídos na tabela 4.16. Ainda com base na amostra de treinamento, a probabilidade da classe de retenção $P(y)$ está definida:

$$P(\text{RET} = \text{“SIM”}) = 0,7669$$

$$P(\text{RET} = \text{“NAO”}) = 0,2331$$

Tabela 4.16: Modelo Geral: Sumarização das probabilidades condicionais dos atributos para as classes de retenção.

Classe: Retenção positiva		Classe: Retenção Negativa	
SENTENÇA: X y	P(X y)	SENTENÇA: X y	P(X y)
ALT_CONTRATUAL = SIM RET = SIM	0,4939	ALT_CONTRATUAL = SIM RET = NÃO	0,4764
ALT_CONTRATUAL = NÃO RET = SIM	0,5061	ALT_CONTRATUAL = NÃO RET = NÃO	0,5236
COD_PRODUTO = 1 RET = SIM	0,2256	COD_PRODUTO = 1 RET = NÃO	0,2072
COD_PRODUTO = 2 RET = SIM	0,3514	COD_PRODUTO = 2 RET = NÃO	0,3804
COD_PRODUTO = 3 RET = SIM	0,1334	COD_PRODUTO = 3 RET = NÃO	0,1413
COD_PRODUTO = 4 RET = SIM	0,2895	COD_PRODUTO = 4 RET = NÃO	0,2711
RESTITUICAO = NÃO RET = SIM	0,9468	RESTITUICAO = NÃO RET = NÃO	0,8967
RESTITUICAO = SIM RET = SIM	0,0532	RESTITUICAO = SIM RET = NÃO	0,1033
ANO_INC_CONT = 1990 RET = SIM	0,002	ANO_INC_CONT = 1990 RET = NÃO	0,0009
ANO_INC_CONT = 1991 RET = SIM	0,0029	ANO_INC_CONT = 1991 RET = NÃO	0,0012
ANO_INC_CONT = 1992 RET = SIM	0,0034	ANO_INC_CONT = 1992 RET = NÃO	0,0016
ANO_INC_CONT = 1993 RET = SIM	0,0071	ANO_INC_CONT = 1993 RET = NÃO	0,0039
ANO_INC_CONT = 1994 RET = SIM	0,0141	ANO_INC_CONT = 1994 RET = NÃO	0,0068
ANO_INC_CONT = 1995 RET = SIM	0,0292	ANO_INC_CONT = 1995 RET = NÃO	0,0176
ANO_INC_CONT = 1996 RET = SIM	0,0417	ANO_INC_CONT = 1996 RET = NÃO	0,0324
ANO_INC_CONT = 1997 RET = SIM	0,0658	ANO_INC_CONT = 1997 RET = NÃO	0,0558
ANO_INC_CONT = 1998 RET = SIM	0,0608	ANO_INC_CONT = 1998 RET = NÃO	0,0581
ANO_INC_CONT = 1999 RET = SIM	0,0583	ANO_INC_CONT = 1999 RET = NÃO	0,0559
ANO_INC_CONT = 2000 RET = SIM	0,0780	ANO_INC_CONT = 2000 RET = NÃO	0,0735
ANO_INC_CONT = 2001 RET = SIM	0,1036	ANO_INC_CONT = 2001 RET = NÃO	0,0938
ANO_INC_CONT = 2002 RET = SIM	0,1116	ANO_INC_CONT = 2002 RET = NÃO	0,0934
ANO_INC_CONT = 2003 RET = SIM	0,1062	ANO_INC_CONT = 2003 RET = NÃO	0,0906
ANO_INC_CONT = 2004 RET = SIM	0,1304	ANO_INC_CONT = 2004 RET = NÃO	0,1071
ANO_INC_CONT = 2005 RET = SIM	0,1078	ANO_INC_CONT = 2005 RET = NÃO	0,1111
ANO_INC_CONT = 2006 RET = SIM	0,0722	ANO_INC_CONT = 2006 RET = NÃO	0,1638
ANO_INC_CONT = 2007 RET = SIM	0,0047	ANO_INC_CONT = 2007 RET = NÃO	0,0325
FAIXA_IDD = 18 - 25 RET = SIM	0,0069	FAIXA_IDD = 18 - 25 RET = NÃO	0,0411
FAIXA_IDD = 26 - 33 RET = SIM	0,1553	FAIXA_IDD = 26 - 33 RET = NÃO	0,2551
FAIXA_IDD = 34 - 41 RET = SIM	0,2006	FAIXA_IDD = 34 - 41 RET = NÃO	0,2191
FAIXA_IDD = 42 - 49 RET = SIM	0,2162	FAIXA_IDD = 42 - 49 RET = NÃO	0,1859
FAIXA_IDD = 50 - 57 RET = SIM	0,1834	FAIXA_IDD = 50 - 57 RET = NÃO	0,158
FAIXA_IDD = 58 - 65 RET = SIM	0,1307	FAIXA_IDD = 58 - 65 RET = NÃO	0,0889
FAIXA_IDD > 65 RET = SIM	0,1069	FAIXA_IDD > 65 RET = NÃO	0,0519

A organização das informações referentes ao modelo de aprendizagem, demonstrados na tabela 4.16, têm como objetivo exibir de forma centralizada os resultados apurados para cada valor de atributo considerado na análise de dados. Para modelagem da base de conhecimento, seria mais

adequado considerar a implementação de um repositório de informações utilizando os conceitos da modelagem entidade-relacionamento. De fato, um repositório centralizado pode ser uma opção viável para concentrar os resultados apurados. Por outro lado, o comportamento dos processos de negócio modificam-se e novas análises podem ser necessárias em um curto espaço de tempo e conseqüentemente, uma revisão dos valores precisará ocorrer. Uma outra vantagem em relação a um modelo entidade-relacionamento está na otimização das consultas e recuperação de informações para grandes lotes de registros.

Os sistemas de suporte à decisão podem utilizar os cálculos de probabilidades condicionais para classificar registros, como forma de prever clientes propensos interromper seu relacionamento com a empresa, com base em algumas características contratuais.

4.2 - Aplicação dos modelos descritivos em estudos de segmentos de clientes

A análise de dados baseada em segmentação também pode ser considerada uma abordagem de classificação de registros, uma vez que gera um modelo descritivo a partir de uma base de dados. O diferencial dessa abordagem é que os segmentos geram modelos mais específicos, em torno de um valor em que se deseja observar para a classe. Sendo assim, se o objetivo é identificar o segmento que descreve a classe negativa de retenção, serão isolados justamente os registros baseados no valor negativo da classe.

As técnicas mais comuns para o estudo de segmentos são a árvore de decisão e a classificação baseada em regras. Por serem relativamente simples na representação do conhecimento e apresentarem bons desempenhos em modelos de aprendizagem, estas técnicas têm sido amplamente utilizados como recursos de suporte à decisão. Um modelo de aprendizagem baseado em segmentos contribui para ações de vendas específicas, como forma de atingir

um determinado público em que já se observou um retorno acima da média quando confrontados com outros grupos de uma mesma ação. Neste trabalho, a modelagem por segmentação é uma proposta de reutilização da base de treinamento para identificação de *Regras de Classificação* e dessa forma, a análise é uma extensão do modelo descritivo principal e representa uma alternativa de aplicação da metodologia de extração de conhecimento.

Para a geração da base de treinamento, utilizou-se o mesmo fluxo de extração de dados, incluindo uma unidade de transformação de dados para identificar os registros. Neste caso, caracterizou-se a "Aquisição de um cliente", cada registro foi novamente classificado em duas categorias de aquisição - Inclusão ou Retorno. Conforme discutido na seção 3. 4, cada regra é composta por dois termos principais: o termo precedente e o termo conseqüente, apresentando o seguinte formato:

$$R : A \rightarrow y$$

onde A é o termo que precede a classe y, sendo composta de um ou mais atributos e respectivos valores.

As regras foram compostas por atributos conforme o índice de cobertura do valor do atributo em relação à classe de aquisição. A partir da inclusão do segundo atributo, avaliou-se o índice de cobertura conjunto para a classe de aquisição, sendo

$$\text{Cobertura}(R) = | A | / | D |$$

onde | A | é a quantidade de registros onde se identificou a regra e | D | a quantidade de registros da amostra. Abaixo, algumas considerações sobre as regras de classificação identificadas são apresentadas.

$$R1: \text{Estado_Civil} = \text{casado} \text{ E } \text{Uf_Residencia} = \text{SP} \rightarrow \text{AQS} = \text{Inclusão}$$

Na regra R1, a cobertura da regra é de $C = 0,4579$ e a precisão $P = 0,5926$. Isso significa que observa-se em aproximadamente 45,79 % dos registros o precedente da regra – “*Estado_Civil = casado E Uf_Residencia = SP*” e dentre os registros que atendem ao precedente da regra – 59,26% são classificados com classe de aquisição correspondente à inclusão de cliente.

Alterando-se um termo da regra R1, com *Uf_Residencia = RJ*, deriva-se a regra R2:

R2: *Estado_Civil = casado E Uf_Residencia = RJ* → AQS = Inclusão

e haverá 1114 registros cobertos pelo precedente da regra R2 e 664 registros para a classe de aquisição = Inclusão correspondentes a essa regra. Embora apresente precisão superior à regra R1, R2 apresenta uma cobertura menor, isto é, abrange apenas 3,168% dos registros.

Ao precedente da regra R1 acrescentou-se mais um termo, gerando a regra R3:

R3: *Estado_Civil = casado E Uf_Residencia = SP E Sexo = M* → AQS = Inclusão

melhorou a precisão da regra, fazendo com que a mesma atingisse $P = 0,6092$, porém a abrangência diminuiu consideravelmente, com $C = 0,3202$.

Considerando um dos objetivos das ações de vendas, a utilização de segmentos tende à especialização das campanhas de vendas, uma vez que cada público tende a uma resposta diferente de acordo com a forma de abordagem e canal de contato utilizado. Uma empresa que faz uma abordagem única para 50000 clientes com dois canais de contato poderá flexibilizar suas campanhas, tendo grupos com diferentes percepções sobre a compra de produtos. Como resultado, serão abordados de forma diferenciada, envolvendo

outros canais de contato. Um possível cenário para uma ação de vendas pode ter origem em uma regra de classificação, conforme descrito na regra 5, em que o termo precedente cobre 353 registros e dentre esses registros, 284 são classificações positivas para a classe "RETORNO".

R5: SE Idade < 35 anos E Uf_Residencia = SP E Sexo = Masculino e Produto_anterior = 1 → AQS = RETORNO

No caso específico da regra R5, haverá uma cobertura baixa se for avaliado o total de registros da categoria de aquisição = "RETORNO". Por outro lado, trata-se de um grupo com alta precisão para recuperação, de aproximadamente 80,45%. Posteriormente, uma abordagem para recuperação dos clientes desse grupo pode ser realizada ressaltando algumas características implícitas na regra, uma vez que são homens e jovens, residentes no estado de São Paulo e já foram clientes do pacote de serviços 1.

As regras de classificação geradas dentro dos procedimentos de extração descritas neste capítulo, podem incorporar bases de conhecimento e posteriormente serem utilizadas por ferramentas de suporte à decisão. Neste caso, é importante que a fase de pós-processamento contenha atividades de modelagem da base de conhecimento ou atualização de bases já existentes. As regras que definem as classes de aquisição de clientes podem ser modeladas em tabelas entidade-relacionamento, conforme figura 4.6. Um sistema de vendas com a funcionalidade de selecionar grupos de clientes para uma determinada campanha pode acessar a base de regras para uma classe de aquisição, e uma vez definida essa classe, o sistema pode recuperar as regras conforme a precisão em relação à classe. Com as regras, o sistema pode acessar os termos precedentes da regra e iniciar a busca na base de dados por clientes que mantém as características definidas na regra.

Figura 4.6: Modelo entidade-relacionamento para modelo de classificação baseado em regras.

4.3 - Otimização do processo KDD

No cenário apresentado, uma empresa que promove ações de retenção de clientes normalmente o faz utilizando poucos recursos de suporte à decisão e na maioria das vezes, envolvendo todos os clientes que estejam próximos da finalização de seus contratos atuais, sem considerar os elementos que caracterizam esses grupos de clientes, tais como rentabilidade e risco de abandono. A implementação de processos de extração de conhecimento em bases de dados permite que tais empresas mantenham procedimentos sistemáticos de análise de dados como alternativa para composição de grupos alvo para ações de vendas a partir de modelos de classificação disponíveis em bases de conhecimento.

Em relação ao ambiente de dados, o cenário ideal para a implementação de um processo de extração de conhecimento é a existência de um ambiente centralizado, com dados históricos da área de domínio sendo constantemente atualizados. Embora esse cenário seja uma realidade para

grande parte das instituições que mantém áreas específicas de planejamento e portanto, usuários potenciais desse ambiente de dados, apenas parte dos sistemas transacionais passam a alimentar um repositório centralizado.

Inicialmente em função da necessidade de uma amostra fechada, estável, onde os dados estarão configurados de tal forma que as consultas poderão ser refeitas sem implicar na extração de casos não envolvidos na extração anterior, implicou na codificação de um programa específico de extração. Neste programa, as extrações de dados foram executadas e as regras de restrição de dados foram incorporadas para eliminar dados inconsistentes ou definir faixas de valores de interesse e níveis de granularidade dos objetos de dados. Abaixo, destaca-se a sequencia de procedimentos em cada etapa do processo KDD para o estudo de retenção de clientes:

- Pré-Processamento.

- (1) Estudo do problema de retenção, bases de dados e atributos disponíveis;
- (2) Escolha do tipo de análise, técnica de mineração de dados;
- (3) Seleção de atributos – implementação de programa de extração – versão inicial, processamento do programa de extração;
- (4) Análise do formato de dados, definição das transformações de dados e criação de faixas de valores, inclusão de regras de restrição para missing values, verificação da cobertura dos valores;
- (5) Alteração do programa de extração, processamento do programa de extração
- (6) Distribuição das bases de dados em amostras de treinamento e amostras de teste.

- Mineração de Dados e Pós- Processamento:

- (1) Isolamento das classes de retenção;
- (2) Cálculo das probabilidades condicionais para as classes de retenção

- (3) Criação do modelo de aprendizagem em modelo de dados entidade-relacionamento e criação das tabelas em banco de dados: uma tabela por atributo, com chave composta por classe de retenção e valor do atributo.
- (4) Criação da estrutura física para teste do modelo de classificação, associando base de teste às tabelas do modelo de aprendizagem
- (5) Processamento da rotina classificação: codificação da extração das probabilidades condicionais para cada valor de atributo da base de teste e codificação do Naïve Bayes.
- (6) Apuração dos resultados da classificação usando o modelo de aprendizagem.

Nos dois problemas apresentados neste capítulo, o produto final disponibilizado pelas técnicas de mineração de dados, apresentam-se em formatos diferentes. O modelo de aprendizagem descritivo, com o cálculo das probabilidades condicionais por classe, pode ser incorporado e mantido em tabelas de bancos de dados relacionais e acessado por sistemas especialistas que implementam o algoritmo de Naïve Bayes para classificação de novos exemplos. Em relação ao modelo de regras de classificação, as unidades de conhecimento representadas por regras e respectivas classes, podem igualmente ser armazenadas em tabelas, mas a dinâmica de sua utilização difere em relação ao modelo descritivo, pois as regras de classificação tendem a selecionar um grupo por meio de uma regra.

5. Conclusões e perspectivas futuras

Este trabalho apresenta uma proposta de modelo de processo de extração de conhecimento, em que a aquisição de informações para a análise de dados têm como origem os bancos de dados transacionais e *data warehouse*.

No capítulo 2, formalizou-se um modelo de processo KDD, com três etapas principais: pré-processamento, mineração de dados e pós-processamento, a partir da análise e unificação de modelos de processo KDD disponíveis. Esse capítulo apresentou ainda os elementos essenciais da fase de pré-processamento, abordando a importância da análise da cadeia de atividades organizacionais para entendimento das bases de dados. Tal abordagem tornou-se necessária pois os bancos de dados mantêm a distribuição física das informações a partir de interpretações das atividades organizacionais aos modelos de dados dos sistemas transacionais. Uma vez que o processo organizacional tenha seu fluxo de trabalho atendido por sistemas transacionais, pode-se assumir que ocorre um mapeamento das informações geradas por esses sistemas a repositórios de dados físicos. Desta forma, identificou-se a

necessidade de explorar tais processos geradores de informação antes de analisar as informações disponíveis nos bancos de dados, como forma de identificar objetos de dados e suas transformações ao longo de uma atividade organizacional.

O capítulo 3 abordou os termos e conceitos da mineração de dados, com foco na geração de modelos descritivos a partir de técnicas de classificação baseadas no Teorema de Bayes e o método direto de extração de regras de classificação. Com essas duas técnicas, definiu-se uma metodologia para a geração de modelos de aprendizagem, os quais permitem descrever, em termos de classes, os registros de um banco de dados, e ainda de se utilizar o modelo de aprendizagem para classificações de novos registros. Como atividade de pós-processamento, foi proposto que a avaliação do modelo de aprendizagem fosse feita com base na apuração das classificações realizadas usando bases de teste.

O capítulo 4 apresentou uma implementação de processo de extração de conhecimento para a geração de modelos de aprendizagem para suporte à decisão, aplicando técnicas de mineração de dados para modelos descritivos e geração de regras de classificação. Explorou-se a possibilidade de transformar os modelos de aprendizagem em bases de conhecimento utilizando um banco de dados relacional, disponível para acesso via sistema especialista, para a realização de novas classificações de registros, ou então para a visualização dos resultados a partir de planilhas eletrônicas. No cenário descrito no capítulo 4, a organização dos procedimentos da etapa de pré-processamento permitiu que a extração de atributos adicionais ou transformação de dados fosse realizada de forma iterativa, sem a necessidade de implementação de novos programas de extração de dados. Isso possibilitou a definição de todas as atividades essenciais do pré-processamento e a sequência em que devem ser realizadas, e ainda a repetição dos procedimentos sem perdas das unidades codificadas para o processo de extração de dados.

Neste trabalho explorou-se a aplicabilidade do modelo de extração de conhecimento e das técnicas de mineração de dados para o estudo do

comportamento de grupos de clientes em ações de retenção. A análise de dados no modelo proposto permitiu a identificação de indicadores de retenção, os quais foram disponibilizados em modelos de aprendizagem com bom desempenho em relação aos procedimentos adotados anteriormente. Com base nos resultados apurados usando os modelos propostos aqui, identificou-se um cenário consideravelmente evoluído no contexto da análise de dados, onde o planejamento das ações de vendas passou a contar com novos recursos e ferramentas para a organização de planos de vendas, com abordagens específicas conforme o perfil dos grupos de clientes selecionados. O cenário encontrado no início dos estudos agregava aos analistas de negócio a exaustiva tarefa de análise de resultados por meio de relatórios, onde as informações relevantes muitas vezes apresentava-se dispersa. No modelo proposto, a composição das bases de conhecimento teve como um dos fundamentos o acesso aos dados por meio de sistemas especialistas, os quais implementaram um algoritmo de decisão de apoio aos profissionais nos processos decisórios.

A área de extração de conhecimento é relativamente recente e encontrar ferramentas que concentrem todas as etapas do processo de extração, como a integração do conhecimento prévio sobre os processos de negócio, seleção dos atributos adequados para a mineração de dados, escolha da técnica de mineração de dados e a avaliação dos resultados, é algo raro. Considerando as ferramentas disponíveis, visualiza-se um processo KDD em que as etapas ainda sejam realizadas dispondo-se de recursos e ferramentas não integradas. Desta forma, o processo de extração de conhecimento foi tratado como um processo de produção de um produto, identificando etapas, metodologias, pontos críticos, produtos e subprodutos, pessoas envolvidas, ferramentas e requisitos de ambiente em cadeia colaborativa. Explorou-se também a possibilidade de tornar o processo iterativo e incremental à medida que novos modelos de aprendizagem sejam gerados, revisando os modelos anteriores. Em relação ao pré-processamento de dados, algumas atividades dependem essencialmente do conhecimento do especialista em mineração de dados e sua habilidade em relacionar os conceitos da área de negócio nos

modelos de dados, e essa habilidade não é coberta pelas ferramentas disponíveis. Este trabalho apresentou novas ferramentas que diminuem a dependência desses especialistas.

Neste trabalho, o conhecimento extraído foi apresentado em dois formatos diferentes: as regras de classificação e o modelo de decisão baseado em probabilidades condicionais em torno de classes. Embora este trabalho abra um caminho para a extração de conhecimento de forma automatizada, ele ainda não resolve todos os desafios de extração de conhecimento, como flexibilizar a manutenção das bases de conhecimento com modelos flexíveis que permitam a incorporação de novas representações de conhecimento oriundas de outras técnicas de mineração de dados.

Para as corporações que presenciam diariamente o crescimento de suas bases de dados, os processos de extração de conhecimento, serão num futuro próximo, parte essencial dos processos corporativos, como base de sustentação dos processos decisórios das empresas e planejamento de ações estratégicas.

A extração de conhecimento em banco de dados, no contexto de um processo organizacional, permite que a análise de dados seja executada dentro de um esquema de trabalho com o objetivo de gerar modelos de aprendizagem a partir de bases de dados. Tais modelos são essenciais para o desenvolvimento de sistemas de suporte à decisão, e alguns cenários onde o processo KDD proposto neste trabalho pode ser utilizado são descritos abaixo:

- ***Risco no transporte de cargas:*** análise dos registros oriundos de sistemas de monitoramento de transporte de cargas para geração de modelo descritivo de risco de perda de carga e conseqüentemente, revisão do custo do frete e avaliação de risco aceitação do seguro de transporte de cargas.
- ***Acompanhamento de gestantes de alto risco:*** formação de base de dados com registros de gestantes - hábitos e histórico de outras gestações e dados registros em exames de acompanhamento, tais como

freqüência cardíaca fetal e contração uterina. Tal estudo permite a criação de uma base de dados para modelo de predição ao risco de parto prematuro ou aborto.

- ***Risco de inadimplência na concessão de crédito:*** a criação de modelos de aprendizagem para identificação dos contratos com alto risco à inadimplência pode utilizar de diferentes tipos de análise de dados, tais como classificação e associação, e com isso, torna-se essencial a adoção de procedimentos de avaliação de desempenho dos modelos gerados, e pois esses são rapidamente incorporados pelos sistemas de análise de crédito. Além disso, a forte concorrência no setor e o alto risco, exigem constante revisão dos modelos.
- ***Modelo de alocação de viaturas de socorro:*** a geração de modelos descritivos para a fluidez do trânsito nos grandes centros urbanos é essencial para os setores de logística. Estes modelos, incorporados a sistemas de alocação de viaturas de socorro, podem ser utilizados para traçar a melhor rota a ser utilizada na atendimento à vítima, bem como a identificação dos melhores pontos onde as viaturas devem aguardar para prestar atendimento. A dificuldade na criação desses modelos é a particularidade do trânsito em cada município e ainda, a quantidade de cenários a serem analisados, a cidade de São Paulo, por exemplo, mantém regras específicas como o rodízio de veículos e regras no tráfego de caminhões que leva à adoção de vários modelos.

Referências Bibliográficas

- ABDULLAH, M. S; EVANS, A., BENEST, I; PAIGE, R., KIMBLE, C. Modelling Knowledge Based Systems using the eXecutable Modelling Framework (XMF). In: *Conference on Cybernetics and Intelligent Systems*, Singapore, 2004.
- ALCOBÉ, J. R. An incremental algorithm for tree-shaped Bayesian network learning. In: *Proceedings of the 10th European Conference of Artificial Intelligence*. IOS Press, 2002, p. 350-355.
- APTE, C., HONG, S. J. Predicting Equity Returns from Securities Data. In *Advances in Knowledge Discovery and Data Mining*. California, EUA: AAAI Press, 1996, p. 541-560.
- AGRAWAL, R., IMIELINSKI, T., SWAMI, A. Mining association rules between sets of items in large databases. In: *Proceedings of the ACM SIGMOD Int'l Conf. On Management of Data*. Washington, EUA, 1993, p. 207-216.
- BABCOCK, C. Parallel Processing Mines Retail Data. In: *Computer World*, 1994, v. 6.
- BARANAUSKAS, J. A. Extração Automática de Conhecimento por Múltiplos Indutores, São Carlos, SP, Brasil: ICMC-USP, 2001.
- BRACHMAN, R. J., ANAND, T. The process of Knowledge Discovery in Databases: A human centered approach. In: *Advances in Knowledge Discovery and Data Mining*. California, EUA: AAAI Press, 1996.
- BERRY, M. J. A., LINOFF, G. S. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Managenent. New York, NY, EUA: Wiley, 2004.
- CLARK, P., BOSWELL, R. Rule Induction with CN2: Some Recent Improvements. In: *Proc. of The 5th European Conference of Machine Learning*. 1991, p. 151-163.
- CLARK, P., NIBLETT, T. The CN2 Induction Algorithm. In: *Machine Learning*, 1989: v. 3, p. 261-283.
- COHEN, W. W. Fast Effective Rule Induction. In: *Proc. of The 12th International Conference on Machine Learning*. Tahoe City, CA, EUA: 1995, p. 115-123.

- CHAPMAN, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T.; SHEARER, C.; WIRTH, R. CRISP_DM 1.0. Disponível em www.crisp-dm.org. 2000.
- DOMINGOS, P., PAZZANI, M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. In: *Machine Learning*: 1997, v. 29, p. 103-130.
- DUDA, R., HART, P., STORK, D. Pattern Classification. New York, NY, EUA: John Wiley and Sons, 2001.
- DUNHAM, M. H. Data Mining: Introductory and Advanced Topics. New Jersey, USA: Prentice Hall, 2002.
- FASMAN, K., CUTICCHIA, A., KINGSBURY, D. The GDB Human Genome Database Anno 1994. Nucl. Acid. R., 22(17): 3462-3469, 1994.
- FAYYAD, U., DJORGOVSKI, G., WEIR, N. Automating the Analysis and Cataloging of Sky Surveys. In: *Advances in Knowledge Discovery and Data Mining*. California, EUA: AAAI Press, 1996, p. 471-494.
- FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. California, EUA: AAAI Press, 1996, p. 37-54.
- FAYYAD, U., STOLORZ, P. Data Mining and KDD: Promise and Challenges. In: *Future Generation Computer Systems*. Redmond, WA, EUA: Elsevier, 1997, v. 13, p. 99-115.
- FISHER, R. A. The use of multiple measurements in taxonomic problem. In: *Annals of Eugenics*, 1936, v. 7, p. 179-188.
- FRAWLEY, W., PIATETSKY-SHAPIRO, G., MATHEUS, C. Knowledge Discovery in Databases: an overview. AI Magazine, 1992, v.14, p.57-70.
- HAN, J., KAMBER, M. Data Mining: Concepts and Techniques. San Francisco, California, EUA: Morgan Kaufmann Publishers, 2006.
- HAND, D. J. Discrimination and Classification. New York, NY, EUA: Wiley, 1981.
- HAND, D. J., MANNILA, H., SMYTH, P. Principles of Data Mining. Massachusetts, EUA: MIT Press, 2001.

- HARRISON, T. H. Intranet Data Warehouse. São Paulo, SP, Brasil: Berkeley Brasil, 1998.
- HIPP, J., MANGOLD, C., GÜNTZER, U., NAKHAEIZADEH, G. Efficient Rule Retrieval and Postponed Restrict Operations for Association Rule Mining. In: *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKK'02)*. Taipei, Taiwan: Springer, 2002, p. 52-65.
- HIPP, J., GÜNTZER, U., NAKHAEIZADEH, G. Data Mining of Association Rules and the Process of Knowledge Discovery in Databases. In: *Advances in Data Mining in E-Commerce, Medicine and Knowledge Management*. London: Springer, 2002, p. 15-36.
- HOLSHEIMER, M., KERSTEN, M., MANNILA, H., TOIVONEN, H. A perspective on databases and data mining. In: *Proc. Of the First International Conference on Knowledge Discovery and Data Mining*. California, EUA: AAAI Press, 1995, p. 150-155.
- HOSKING, J. R. M., PEDNAULT, E. P. D., SUDAN, M. A statistical perspective on data mining. In: *Future Generation Computer Systems*. New York, NY, EUA: Elsevier, 1997, v. 13, p. 117-134.
- JAIN, A. K., DUBES, R. C. Algorithms for Clustering Data. Prentice Hall, 1988.
- JAIN, A. K., DUIN, R. P. W., MAO, J. Statistical Pattern Recognition: A Review. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n.1, p. 4-37, 2000.
- JENSEN, F., OLSEN, K., ANDERSEN, S. An Algebra Bayesian Belief Universes for Knowledge-based Systems. In: *Networks*. New York, NY, EUA: 1990, v. 20, p.637-659.
- KEOGH, E., PAZZANI, M. Learning augmented bayesian classifiers: A comparison of distribution-based and classification-based approaches. 1999.
- KIMBALL, R., The Data Warehouse Toolkit. New York, NY, EUA: John Wiley, 1996.

- KURGAN, L., MUSILEK, P. A survey of Knowledge Discovery and Data Mining process models. In: *The Knowledge Engineering Review*. Reino Unido: Cambridge University Press, 2006, v. 21, p. 1-24.
- LANGLEY, P., IBA W., THOMPSON, K. An analysis of Bayesian Classifiers. In: *Proc. of the 10th National Conference on Artificial Intelligence*. 1992, p. 223-228.
- LEWIS, D. D. Naïve Bayes at Forty: The Independence Assumption in Information Retrieval. In: *Proc. of The 10th European Conference on Machine Learning*. 1998, p. 4-15.
- LUGER, G. Inteligência Artificial: estruturas e estratégias para a solução de problemas complexos. Traduzido por: Paulo Engel. Porto Alegre, RS, Brasil: Bookmann, 2004.
- McGARRY, K. A survey of interestingness measures for knowledge discovery. In: *The Knowledge Engineering Review*. Reino Unido: Cambridge University Press, 2005. v. 20, p. 39-61.
- McLACHLAN, G. Discriminant Analysis and Statistical Pattern Recognition. New York, NY, EUA: Wiley, 1992.
- MICHALSKI, R. S., MOZETIC, I., HONG, J., LAVRAC, N. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In: *Proc. of 5th National Conference on Artificial Intelligence*. Orlando, EUA: 1986.
- PARR RUD, O. Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management. New York, NY, EUA: John Wiley & Sons, 2001.
- PYLE, D. Business Modeling and Data Mining. San Francisco, CA, USA: Morgan Kaufmann, 2003.
- RAMONI, M., SEBASTIANI, P. Robust Bayes Classifiers. In: *Artificial Intelligence*. 2001, v. 125, p. 209-226.
- SCHREIBER, G., AKKERMANS, H., ANJEWIERDEN, A., de HOOG, R., SHADBOLT, N., de VELDE, W. V., WIELINGA, B. Knowledge Engineering

- and Management: The Common KADS Methodology. Boston, MA, EUA: MIT Press, 1999.
- SPEEL, P. H., SCHREIBER, A. T., JOOLINGEN, W., HEIJST, G., BEIJER, G. J. Conceptual Modelling for Knowledge-Based Systems. In: *Encyclopedia of Computer Science and Technology*. New York, NY, EUA: Marcel Dekker Inc., 2001.
- TAN, P. N., STEINBACH, M., KUMAR, V. Introduction to Data Mining. Boston, MA, EUA: Addison-Wesley, 2006.
- TEOREY, T., LIGHTSTONE, S., NADEAU, T. Database Modeling and Design. San Francisco, CA, EUA: Kaufmann Publishers, 2006.
- ULLMAN, J. Principles of Database and Knowledge-Base Systems. Maryland, EUA: Computer Science Press, 1988.
- YANG, Y. Discretization for Naïve Bayes Learning. PhD thesis, School of Computer Science and Software Engineering of Monash University, 2003.
- YOON, J., KERSCHBERG, L. A framework for knowledge discovery and evolution databases. In: *IEEE Transaction on Knowledge and Data Engineering*, 1993, v. 5, p. 973-979.
- WATANABE, S. Pattern Recognition: Human and Mechanical. New York, NY, EUA: Wiley, 1995.
- WATERMAN, D. A. A Guide to Expert Systems. Canada: Addison-Wesley, 1986.
- WEISS, S., KULIKOWSKI, C. Computer Systems that Learn: Classification and Predictions Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. California, EUA: Morgan Kaufmann Publishers, 1991.
- WILLIAMS, G., HUANG, Z. Modeling the KDD process: A four stage process and four element model. In: *Technical Report TR-DM-96013*, CSIRO Division of Information Technology, Canberra: 1996, p. 1-8.
- WIRTH, R., HIPPE, J. CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*. Manchester: 2000, p. 29-39.

WITTEN, I., FRANK, E. Data Mining: Pratical machine learning tools and techniques with Java implementations. San Diego, EUA: Academic Press, 2005.